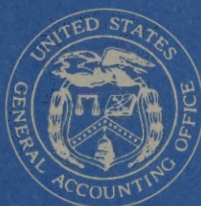


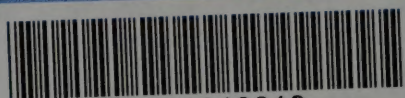
March 1992

# CROSS DESIGN SYNTHESIS

## A New Strategy for Medical Effectiveness Research







22501848819



Program Evaluation and  
Methodology Division

B-244808

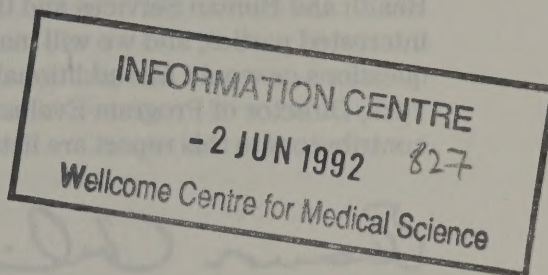
March 17, 1992

The Honorable George J. Mitchell  
Majority Leader  
United States Senate

The Honorable John Glenn  
Chairman, Committee on  
Governmental Affairs  
United States Senate

The Honorable David Pryor  
Chairman, Special Committee  
on Aging  
United States Senate

The Honorable William S. Cohen  
Ranking Minority Member  
Special Committee on Aging  
United States Senate

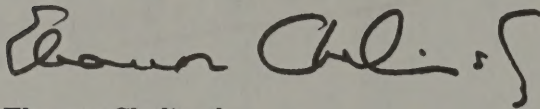


In December 1989, the Congress established the Agency for Health Care Policy and Research (AHCPR) in an effort to improve the quality of health care afforded the American public. As an accompaniment to this action, you asked us to initiate a series of studies that would assist AHCPR with its mission and, most particularly, with determining the effectiveness of medical interventions.

We have already provided you with the results of two of these studies: (1) an examination of the obstacles AHCPR is likely to face in obtaining data on effectiveness and (2) a review of the methods medical specialty societies have used in the past to develop medical practice guidelines, based on whatever was known about treatment effectiveness. In this report, we assess ways to determine how well medical treatments actually work, and we present a new strategy for medical effectiveness research. This strategy, which we call "cross design synthesis," represents an important step in advancing knowledge about the effectiveness of medical treatments, based not only on the results of randomized clinical trials but also on data reflecting wide experience in the field.

---

As agreed with your offices, we will send copies of this report to officials of the Department of Health and Human Services and the Agency for Health Care Policy and Research and to other interested parties, and we will make copies available to others upon request. If you have any questions or would like additional information, please call me at (202) 275-1854 or Robert L. York, Director of Program Evaluation in Human Service Areas, at (202) 275-5885. Major contributors to this report are listed in appendix II.

A handwritten signature in cursive script, appearing to read "Eleanor Chelimsky".

Eleanor Chelimsky  
Assistant Comptroller General



Findings

GAO's review of the literature on the effectiveness of health care policy and research found that the literature is fragmented and often contradictory. The literature is often limited to a specific aspect of health care, such as the effectiveness of a specific intervention, rather than providing a comprehensive overview of the health care system. The literature also tends to be dated, with many studies published more than 10 years ago. This limits the applicability of the findings to current health care practice. The literature also tends to be biased, with many studies funded by the pharmaceutical industry or other interested parties. This can lead to a lack of objectivity and a focus on positive results. The literature also tends to be fragmented, with many studies focusing on a specific aspect of health care, rather than providing a comprehensive overview of the health care system. This makes it difficult to synthesize the findings and draw conclusions about the effectiveness of health care policy and research. The literature also tends to be biased, with many studies funded by the pharmaceutical industry or other interested parties. This can lead to a lack of objectivity and a focus on positive results. The literature also tends to be fragmented, with many studies focusing on a specific aspect of health care, rather than providing a comprehensive overview of the health care system. This makes it difficult to synthesize the findings and draw conclusions about the effectiveness of health care policy and research.

Background

There is a significant gap in the literature on the effectiveness of health care policy and research. The literature is often limited to a specific aspect of health care, such as the effectiveness of a specific intervention, rather than providing a comprehensive overview of the health care system. The literature also tends to be dated, with many studies published more than 10 years ago. This limits the applicability of the findings to current health care practice. The literature also tends to be biased, with many studies funded by the pharmaceutical industry or other interested parties. This can lead to a lack of objectivity and a focus on positive results. The literature also tends to be fragmented, with many studies focusing on a specific aspect of health care, rather than providing a comprehensive overview of the health care system. This makes it difficult to synthesize the findings and draw conclusions about the effectiveness of health care policy and research. The literature also tends to be biased, with many studies funded by the pharmaceutical industry or other interested parties. This can lead to a lack of objectivity and a focus on positive results. The literature also tends to be fragmented, with many studies focusing on a specific aspect of health care, rather than providing a comprehensive overview of the health care system. This makes it difficult to synthesize the findings and draw conclusions about the effectiveness of health care policy and research.

Results in Brief

GAO conducted a critical review of the literature on the effectiveness of health care policy and research. The review found that the literature is fragmented and often contradictory. The literature is often limited to a specific aspect of health care, such as the effectiveness of a specific intervention, rather than providing a comprehensive overview of the health care system. The literature also tends to be dated, with many studies published more than 10 years ago. This limits the applicability of the findings to current health care practice. The literature also tends to be biased, with many studies funded by the pharmaceutical industry or other interested parties. This can lead to a lack of objectivity and a focus on positive results. The literature also tends to be fragmented, with many studies focusing on a specific aspect of health care, rather than providing a comprehensive overview of the health care system. This makes it difficult to synthesize the findings and draw conclusions about the effectiveness of health care policy and research. The literature also tends to be biased, with many studies funded by the pharmaceutical industry or other interested parties. This can lead to a lack of objectivity and a focus on positive results. The literature also tends to be fragmented, with many studies focusing on a specific aspect of health care, rather than providing a comprehensive overview of the health care system. This makes it difficult to synthesize the findings and draw conclusions about the effectiveness of health care policy and research.



# Executive Summary

---

## Purpose

With the establishment of the Agency for Health Care Policy and Research (AHCPR) in December 1989, the Congress launched an "effectiveness initiative" intended to improve the quality of health care through research-based development of national guidelines for medical practice. Credible guidelines require research results that are both scientifically valid and relevant to the conditions of medical practice. However, there is an increasing realization of just how elusive such information can be, and the task facing the new agency is a difficult one. Thus, four Senators asked GAO to (1) review existing designs for evaluating medical effectiveness, and (2) suggest an evaluation strategy that avoids the limitations of existing approaches.

---

## Background

There is a surprising lack of knowledge about "what works in medicine." Many technologies in common use have never been evaluated. Many of those that have been evaluated remain of uncertain benefit to large numbers of patients. This is true even when the evaluations were conducted in accordance with the highest scientific standards.

The problem derives from the fact that controlled studies are typically conducted under conditions much more limited than those in which medical practice occurs. In actual medical practice, what works well for one type of patient may be less effective for others. A new technique that is highly effective when implemented by an expert may prove less so in the hands of a relatively inexperienced medical team. And the treatment that improves patient survival may have surprisingly negative effects on other outcomes, such as quality of life.

Scientific study designs are not suited to capturing the range of relevant patients, treatment implementations, and outcome criteria that count in medical practice. New research strategies are needed to provide the broader knowledge that is essential for setting practice guidelines. The goal is to achieve a research base that is at once scientifically sound and relevant for real-life patients and physicians across the United States.

---

## Results in Brief

GAO conducted a critical review of study designs that have been used to evaluate how well medical treatments "work." GAO found that all study designs are characterized by strengths and weaknesses. GAO also found that certain combinations of designs are complementary: The chief weakness of one study design may occur in an area where another design is strong. In particular, two study designs were found to have



complementary strengths and weaknesses: randomized studies and data base analyses. Randomized studies can provide the scientific rigor needed to evaluate how well a treatment works, but their results may not be fully generalizable to the varied conditions of medical practice. Data base analyses are potentially weak because patients are not randomly assigned to alternative treatments. “Imbalanced comparison groups” may result. If corrected for this problem, the results of many data base analyses would be highly generalizable to actual medical practice.

GAO’s review of study designs also showed that considerable work has been done in combining research results across existing studies. This work, known as meta-analysis or quantitative “overview,” specifies ways of combining results from similar studies and, to some extent, across more diverse study designs.

Building on findings from the review of designs, GAO devised a strategy that extends the logic of meta-analysis. The new strategy, which GAO terms “cross design synthesis” and presents in this report, combines results from studies that have different, complementary designs. The goal of this strategy is to capture the strengths of the different designs, while minimizing weaknesses. This goal is pursued by capitalizing upon numerous existing techniques for assessing, adjusting, and combining the results of existing studies.

To gauge the feasibility of cross design synthesis, GAO developed a methodology for combining results from two complementary designs—randomized studies and data base analyses. This methodology is designed to answer the following research question:

Does the treatment “work” across the full range of patients for whom it is intended?

Based on the methodology presented here, along with a review of this work by a panel of experts, GAO concludes that cross design synthesis is indeed a feasible strategy.

---

## GAO’s Analysis

Drawing upon established and lesser known techniques from methodological and substantive literature in a number of fields, GAO identified the methods necessary for a synthesis of randomized studies and data base analyses. GAO brought these methods together in a series of



tasks and steps that constitute a first-cut methodology for determining whether a treatment “works” across the full range of patients.

The four major tasks of this methodology are:

- Assess existing randomized studies for generalizability across the full range of relevant patients (task 1).
- Assess data base analyses for “imbalanced comparison groups” (task 2).
- Adjust the results of each randomized study and each data base analysis, compensating for biases as needed (task 3).
- Synthesize the studies’ adjusted results within and across design categories (task 4).

Tasks 1 and 2, above, constitute the cornerstone of the strategy. A cross design synthesis is needed only if assessment shows that randomized studies’ results are not generalizable to the relevant patient population. Further, a cross design synthesis is possible only if assessment shows that data base analyses are sufficiently valid. In-depth assessment also provides the basis for secondary adjustment of each study’s results (task 3). Finally, the assessments provide information that clarifies the range of uncertainty associated with each study’s results. This guides key decisions in task 4.

GAO reviewed and assembled existing techniques for assessing the generalizability of randomized studies (that is, techniques for conducting task 1). These techniques were combined in a set of appropriate steps, ranging from logic-based assessment of how patients were enrolled in these studies to empirical analyses that compare the patients who were enrolled to those who are seen in medical practice.

GAO found that data base analyses are potentially weak, primarily because of the potential for imbalanced comparisons. Techniques for assessing imbalance (task 2) include reviewing the methods used by the primary analyst and conducting a variety of empirical tests.

Adjusting individual study results to compensate for biases (task 3) involves using specific information generated by the assessments to standardize each randomized study’s results to distributions in the patient population (that is, to compensate for known underrepresentation and overrepresentation of key patient groups). The assessment information is also used either (1) to raise or lower each treatment effect estimated by a data base analysis, thus correcting for known imbalances in comparison groups, or (2) to define ranges that account for potential imbalances.



Finally, synthesizing adjusted results from both kinds of studies (task 4) requires a framework that defines appropriate categories of study design and population coverage, the use of meta-analysis techniques to combine study results, and the use of other statistical methods, such as projection, to extend results to patient groups not covered by randomized studies.

As presented, cross design synthesis has three major strengths. First is the capacity to draw upon different kinds of studies that, in combination, can tell more about how medical treatments work than any single type of study can. Second, cross design synthesis can be applied to existing results in several areas because diverse study designs are increasingly being used to evaluate treatment effectiveness. Indeed, when the different designs yield divergent findings, there is a special need for this approach. Third, cross design synthesis is a new and efficient evaluation strategy that can produce the scientifically strong and generalizable information that is needed to develop credible medical practice guidelines. Given these strengths, GAO believes that the further development and use of this research strategy should be facilitated.

The major limitation of cross design synthesis is that it requires investigator judgment for many decisions. Until refinements of this strategy are developed, GAO believes it is best applied by those knowledgeable about both a specific medical treatment and evaluation methods in general.

## Agency Comments

Because this study did not examine AHCPR policies and procedures, GAO did not seek written agency comments. GAO periodically updated agency representatives on this study and received their informal reactions. Throughout, GAO consulted experts in medicine, statistics, meta-analysis and evaluation design; these experts are listed in appendix I.



# Contents

<b>Executive Summary</b>		<b>4</b>
<hr/>		
<b>Chapter 1</b>		<b>10</b>
<b>Introduction and</b>	Three Dimensions of Effectiveness	12
<b>Review of Study</b>	The Scope of Our Study	15
<b>Designs for</b>	Review of Existing Study Designs	17
<b>Evaluating Treatment</b>	Implications of the Review of Designs	27
<b>Effectiveness</b>	Objective, Approach, and Methodology	30
	Organization of the Report	33
<hr/>		
<b>Chapter 2</b>		<b>34</b>
<b>Methods for Assessing</b>	Judgmental Assessments of Patient Selection Methods	35
<b>Randomized Studies'</b>	Empirical Assessments of the Representativeness of the	40
<b>Generalizability to the</b>	Achieved Patient Pool	
<b>Patient Population</b>	Summary of Task 1: Steps in Assessing the Generalizability	49
	of Existing Randomized Studies	
<hr/>		
<b>Chapter 3</b>		<b>52</b>
<b>Methods for Assessing</b>	Judgmental Assessments of Methods Used by Data Base	53
<b>Data Base Analyses</b>	Analysts	
<b>for Comparison Bias</b>	Empirical Assessments of Achieved Balance in Adjusted	64
	Comparison Groups	
	Summary of Task 2: Steps in Assessing Comparison Bias	74
<hr/>		
<b>Chapter 4</b>		<b>77</b>
<b>Methods for Adjusting</b>	Logic of Task 3: Adjusting Individual Studies' Results	77
<b>and Combining</b>	Logic of Task 4: Combining Results Within and Across	79
<b>Results of</b>	Design Categories	
<b>Randomized Studies</b>	Tasks, Steps, and Challenges in Adjusting and Combining	81
<b>and Data Base</b>	Studies	
<b>Analyses</b>	Making Secondary Adjustments to Enhance	82
	Generalizability	
	Making Secondary Adjustments to Minimize Comparison	84
	Bias	
	Designing a Synthesis Framework	86
	Combining Study Results Within Design Categories	88
	Synthesizing Results Across Design Categories	93



	Summary of Tasks 3 and 4: Adjusting and Combining Studies	96
<b>Appendixes</b>	Appendix I: List of Experts	98
	Appendix II: Major Contributors to This Report	99
	Bibliography	100
<b>Tables</b>	Table 1.1: Complementarity of Strengths and Weaknesses of Two Study Designs	29
	Table 2.1: Logic of Patient Outcome Comparisons: Patterns Signaling Nongeneralizability of Randomized Studies	48
	Table 2.2: Assessing Generalizability of Randomized Study Results to the Patient Population: Four Steps	50
	Table 3.1: Assessing the Analyst's Choice of Comparison Groups: Treatment Groups Versus Natural Cohorts	59
	Table 3.2: Logic of Patient Outcome Comparisons: Patterns Signaling Comparison Bias in a Data Base Analysis	73
	Table 3.3: Assessing Imbalanced Comparisons in Data Base Analyses: Five Steps	75
	Table 4.1: Adjusting and Combining Studies: Challenges, Tasks, and Steps	82
	Table 4.2: Synthesis Framework: Primary and Secondary Dimensions of Stratification	88
	Table 4.3: Example of a Plan to Account for Multiple Cross-study Differences in Four Randomized Studies	92
	Table 4.4: Set of Strategies for Adjusting and Combining Diverse Studies	97
<b>Figure</b>	Figure 1.1: The Effectiveness Domain	13

### Abbreviations

AHCPR	Agency for Health Care Policy and Research
AIDS	Acquired immunodeficiency syndrome
APACHE	Acute Physiology and Chronic Health Evaluation
CASS	Coronary Artery Surgery Study
CDC	Centers for Disease Control
GAO	General Accounting Office
HCFA	Health Care Financing Administration
MRFTT	Multiple Risk Factor Intervention Trial
SEER	Surveillance, Epidemiology, and End Results Data Base



# Introduction and Review of Study Designs for Evaluating Treatment Effectiveness

Our nation's "crisis of health care" has been defined mainly in terms of excessively high and ever-escalating cost, but there are also indications that, despite its high cost, the quality of health care is not optimal. The problem of quality has many manifestations, some of which point to a surprising lack of knowledge about "what works in medicine." Indeed, more and better information is needed if the nation is to discriminate among medical interventions that are clearly beneficial for specific types of patients, those interventions that are less so, and those that do little but drive up the costs of health care and may even subject some patients to needless risk.

Many new medical technologies in common use have never been rigorously evaluated (Mosteller et al., 1985). Many other treatments have been evaluated in studies that cannot be generalized across the full range of medical practice. Numerous procedures have been routinely applied to broad classes of patients—even when the benefits of these procedures are uncertain for all patients or are known only for certain kinds of cases.

Electronic fetal monitoring is a case in point. This procedure was introduced for use with high-risk births in the early 1970s and, over time, came to be used routinely for normal deliveries. Sizable scientific evaluations eventually showed that, aside from high-risk cases, electronic fetal monitoring is no more effective than the earlier standard approach—which was based on intermittent use of the stethoscope (see Leveno et al., 1986). Now, the American College of Obstetricians and Gynecologists has concluded that electronic fetal monitoring provides no benefit over the stethoscope (Sandmire, 1990). Hysterectomy and drug therapy for hypertension are two other examples of procedures that, in the absence of comprehensive studies, were applied to much broader patient groups than originally intended (see Mosteller et al., 1985).

These examples suggest that the lack of information about "what works in medicine" has led to the overuse of certain treatments—one source of needless increases in the cost of care. There may also be parallel instances of underutilization, although these seem more difficult to identify.

To improve the quality of health care, the Congress established the Agency for Health Care Policy and Research (AHCPR) in December 1989. One way in which the new agency is to move toward improved quality of care is by conducting rigorous evaluations to produce relevant information on the effectiveness of medical interventions. But there is an increasing realization of how elusive accurate information can be. Leading experts in



medical effectiveness research call not only for the evaluation of more treatments using the strongest evaluation designs, but also for methodological work to strengthen the “weaker” designs (see Mosteller et al., 1985; Sechrest and Hannah, 1990). New strategies are needed because of the high cost and the inherent limitations of the strongest existing designs.

Shortly before the creation of AHCPR, we received a request from Senators Mitchell, Glenn, Pryor, and Heinz to examine current methods for studying medical effectiveness and to suggest new strategies for effectiveness research. This report describes one of the efforts we have mounted in an on-going series responding to this request.<sup>1</sup>

Research on medical effectiveness spans several question areas, including, for example: (1) how well alternative interventions actually “work” in medical practice; (2) how evaluation results are linked to medical practice guidelines; (3) how widely certain treatments are used in community medical practice; and (4) the relative costs of alternative procedures. As implied in the Senators’ request, the first question area—how well alternative interventions actually “work” in medical practice—is central and poses many challenges.

A prior report in this series examined the second question area: the development of practice guidelines (GAO, 1991). As that report made clear, practice guidelines for any one disease or condition should be based on sound information about the relative benefits of the alternative medical interventions that are available. At a minimum, such guidelines should reflect the full degree of uncertainty in our current knowledge about the true benefits that specific treatments provide when they are applied to real patients in actual medical practice. Such considerations underscore the importance and primacy of the first question area, which we address in this report.

Specifically, in accordance with the Senators’ request, this report (1) reviews existing designs for evaluating how well medical interventions work, and (2) suggests an evaluation strategy that improves upon existing research approaches for answering this question.

The scope of this study is limited in two important ways. First, we limited the review and development of study designs to those suitable for

---

<sup>1</sup>Earlier studies examined (1) the relevance of data base information to key questions in medical effectiveness research (results reported in a briefing of committee staff) and (2) the experience of medical societies in developing practice guidelines (see GAO, 1991).



evaluating treatment effects, thus excluding diagnostic procedures. Second, in developing an improved evaluation strategy, we focused on achieving both (1) scientific rigor, and (2) generalizability of results across all patients for whom the treatment will be used in medical practice. As explained below, our focus on extending results across all patients is coupled with both the assumption of a constant treatment implementation and the assumption of a single, objective outcome criterion. Future work is planned to extend the results of this study beyond the limitations of these assumptions.

---

## Three Dimensions of Effectiveness

“Effectiveness” is a commonly used word, which can be defined as the state of accomplishing a desired outcome. “Medical effectiveness” refers to the extent to which treatments accomplish certain outcomes. For example, the effectiveness of aspirin for headaches might be determined by the degree to which aspirin alleviates the pain. Similarly, the effectiveness of Band-Aids for cuts would be indicated by whether they keep the wound clean, allow it to heal properly, and so forth. However, despite the clarity of these simple examples, medical effectiveness is a complex concept.

The “effectiveness domain” involves three major dimensions or sources of complexity: (1) various types of patients and forms of the disease, (2) varying implementations of the treatment in question, and (3) varying outcome criteria.

- A treatment may be more effective for certain kinds of patients than for others. For example, a given dosage of aspirin may be effective in helping patients with moderate headaches, but less likely to ease pain for those with severe migraines.
- A treatment may be less effective if it is implemented in a less than optimal fashion. For example, the process by which Band-Aids help minor cuts is not mysterious; however, a study might show that, in many instances, Band-Aids did not protect cuts from either dirt or abrasion because they were put on haphazardly by children and either fell off quickly or never adequately covered the cut in the first place. As this example shows, there is a difference between the potential benefit of a medical intervention and the realization of that potential.<sup>2</sup>
- A treatment may appear to be more or less effective depending upon the particular type of outcome measure that is used as a criterion. For

---

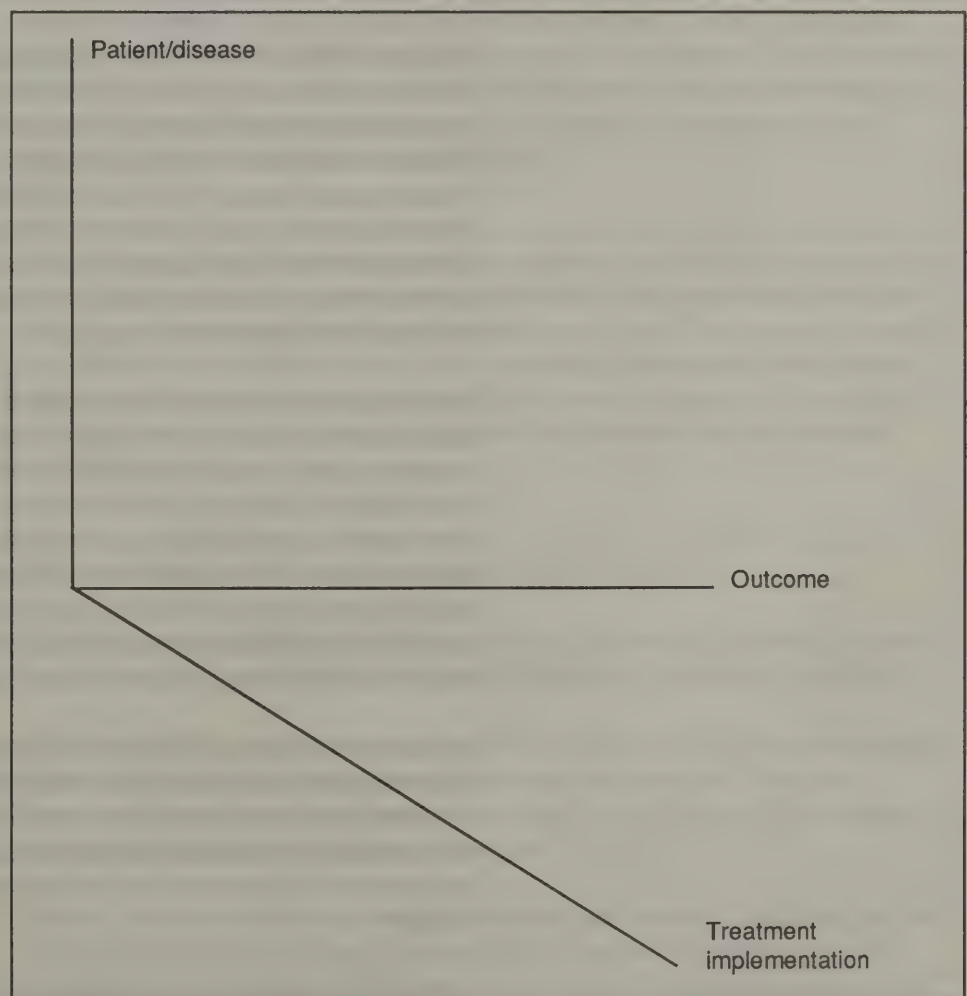
<sup>2</sup>Medical effectiveness has been defined as incorporating assessments of both a treatment's potential (generally referred to as “efficacy”) and the realization of that potential.



example, aspirin might appear to “work well” if the outcome measured was pain reduced (or not) within 15 minutes. But it might appear less effective on a criterion of all pain eliminated within 5 minutes.

Taken together, the three dimensions shown in figure 1.1 define what we term the “effectiveness domain.”

Figure 1.1: The Effectiveness Domain



The effectiveness of any one treatment (for example, aspirin) may be different for different points within this domain. That is, the precise effectiveness of a treatment (such as aspirin) depends upon (1) the



specific type of patient to whom it is administered (for example, young, otherwise healthy patients with severe migraine headaches), (2) the specific form or level of the treatment that is administered (for example, 900 mg of aspirin), and (3) the specific outcome measure deemed most salient (perhaps whether pain is at least reduced somewhat in 15 minutes).<sup>3</sup> For any one point in the domain, a patient with specific characteristics receives a treatment in a specific form, and a specific type of outcome is measured.

The size of the effectiveness domain is determined, first, by the realities of medical practice; that is, whether the patients who need the treatment in question are a varied group, whether the treatment is implemented in different ways, and whether different kinds of outcomes really do matter. Thus, depending upon the particular treatment in question, one or more dimensions may not be relevant. For example, a centrally produced drug administered in hospitals at a precise dosage would not vary substantially on the treatment implementation dimension. In other instances, all three dimensions are potentially relevant.

Second, given the realities of medical practice, the size of the domain is limited by the robustness of the treatment effect. That is, a very robust treatment (such as penicillin for common infections) likely affects even very different kinds of patients in the same way. The benefits of such a treatment may not depend, in a substantial degree, on exactly how it is implemented. And whether or not such a treatment appears to help patients probably does not depend upon the specific outcome criterion used. But the effectiveness of other, less robust treatments can vary across one, two, or all three of these dimensions.

For many treatments, the size of the effectiveness domain (that is, the dimensions of medical practice and the robustness of the treatment across these dimensions) is unknown. And if available information covers only certain kinds of patients, implementations, and outcome measures, it is not possible to determine how consistent or robust a treatment's effect really is.

Thus, a study that captures only a very limited number of points may not, by itself, adequately capture the full story or "truth" about the effectiveness of the treatment in question. If a rigorous scientific study includes only certain kinds of patients, only selected—perhaps

---

<sup>3</sup>Interactions between dimensions are possible. For example, a treatment may be equally effective in improving quality of life for all patient age groups, but improve survival only for younger patients (see Zelen and Gelman, 1986).



optimal—forms of implementing the treatment, and a single outcome measure, that study can only tell a small part of the story. Such a study may be a poor guide for making many of the decisions required in medical practice. Certainly, an equally scientific study of the same treatment could yield very different results if it focused on different kinds of patients, implementations, and outcomes.

Ideally, information that serves as a basis for setting medical practice guidelines would cover the complete effectiveness domain. Scientific studies have often omitted large portions of that domain. Thus, our review of study designs and our proposal for an improved evaluation strategy is aimed at feasible approaches for achieving greater coverage of that domain, while maximizing scientific rigor.

---

## The Scope of Our Study

In developing the proposed evaluation strategy, we decided to target the simultaneous achievement of scientific validity and extended coverage with respect to one of the three dimensions of the effectiveness domain: the patient/disease dimension. Thus, our study focuses on designs and strategies for evaluating treatment effectiveness across the full range of patients, assuming a constant treatment implementation and a single outcome criterion.

---

## Extending Patient Coverage

The primary effectiveness question we deal with is:

- What is the average effect of the treatment, calculated across all patients for whom it is intended (or to whom it is applied in medical practice)?

Cronbach (1982, p. 155) has emphasized that if a general treatment policy is selected on the basis of “what works best on average,” then the “average” for patients participating in a study should match the average for all patients to whom the general policy will apply.

Related questions concern how effective the treatment is with respect to different patient subgroups. For example, does the treatment provide more or less benefit for patients with different severity of disease or for patients in different demographic groups? Such questions are important because, to the extent that different kinds of patients (for example, the elderly or the most severely ill) respond to a treatment differently than other kinds of patients do, this information should be given to clinicians

(or incorporated into guidelines) so that each patient can receive the most effective individual care.

---

## Simplifying Assumptions

As noted above, our decision to focus on extending coverage of scientific results across the patient dimension (that is, to develop a strategy for studying how well a treatment works across all relevant patients) is coupled with two important assumptions about the other two dimensions of the effectiveness domain.

- First, we assumed that there is a constant implementation of the treatment, as would occur, for example, for a centrally produced drug administered by hospital personnel in a set dosage.
- Second, we assumed that a single, relatively objective outcome measure is of interest; for example, 5-year survival.

The decision to focus on a single dimension in this initial study was made in light of the complex issues relevant to each of the three dimensions. Indeed, the foregoing discussion of the effectiveness domain, though instructive, represents merely the “tip of the iceberg.” For example, the treatment implementation dimension involves not only such issues as surgical skill, dosage level, and so forth, but also the settings in which treatments are given and even the kind of information that is given to patients (which may have psychological effects<sup>4</sup>). Issues associated with the outcome criterion dimension have involved the varying subjectivity-objectivity of different measures and the need for study designs to incorporate the blinded assessment of outcomes (especially for subjective measures; see T.C. Chalmers et al., 1981).

Our decision to focus on extending coverage across the patient/disease dimension in no way implies a lesser importance of extending coverage across the treatment implementation or the outcome criterion dimensions. These need to be examined in future work.<sup>5</sup>

---

<sup>4</sup>The importance of the different types of information about a treatment that might be given to patients is suggested by the “placebo effect”; that is, the effect of an inactive substance that patients are told may or may not be an active drug (see, e.g., Beecher, 1955; Rosenthal, 1985; Kirsch and Weixel, 1988; Wilkins, 1985). Patients enrolled in placebo studies are provided with information that is different from what would be given them in medical practice; for example, those receiving the “real” medicine are sometimes told they may or may not be receiving a placebo.

<sup>5</sup>As previously noted, the most important next steps for future work may involve one or both of these dimensions. For example, a future study might examine whether a methodology parallel to the steps of cross design synthesis proposed in this report can be used to extend the coverage of effectiveness information across the varied treatment implementations that occur in actual medical practice.



---

## Review of Existing Study Designs

Through the years, a variety of studies—with different designs—have been used to evaluate treatment effects. Some study designs have been deemed superior to others; each design, however, has characteristic strengths and weaknesses. In this review, we discuss early approaches, randomized studies, meta-analyses, and data base analyses.

---

## Early Approaches

Considering that the practice of medicine dates back to antiquity, it is only relatively recently (that is, within the past 150 to 200 years) that studies of treatment effectiveness have moved beyond personal observations by individual clinicians. The personal-observation approach relies on the expertise and acumen of the clinician to interpret the observed consequences of treating a specific patient with a particular therapy. The strength of this approach is that it incorporates clinically relevant experiences into the conclusions and thus provides richness to them.<sup>6</sup> But the weaknesses of personal observation are many and include the possibility that the patient outcomes observed are coincidental to the treatment rather than caused by it.

Perhaps the earliest instance of a more objective approach is Lind's 18th century research on treatments for scurvy conducted on board the Salisbury at sea (see Pocock, 1983, citing Lind, 1753). Lind took 12 very similar cases and assigned two each to six different treatments, finding that the treatment consisting of oranges and lemons produced the best result.

P. Louis (1834, 1835) made explicit the limitations of the usual subjective approach and developed the "numerical method" in 19th century France.<sup>7</sup> The numerical method emphasized exact observations, exact recordings of treatments (and deviations from intended treatment), and numerical comparisons of patient outcomes.

Since that time, a variety of controlled designs have been developed. For example, historical control trials compare outcomes for patients currently receiving a new treatment to historically recorded outcomes for patients who, at earlier times, had received different treatments. Such studies apparently provided the evidence supporting the majority of cancer

---

<sup>6</sup>Indeed, case studies are still routinely reported in the *New England Journal of Medicine*, and many medical decisions are probably still made on the basis of individual physicians "trying out" a new drug or a new diagnostic test to "see if it works" (Sechrest and Figueredo, 1991, p. 7).

<sup>7</sup>It was Louis' use of this method that led to a "decline in bleeding as a standard treatment" for many illnesses (Pocock, 1983, p. 15, citing P. Louis, 1835).

treatments in use today (Miké, 1982). Logically, however, the difference between the average outcomes cannot be attributed to the treatments received unless the patients in the current study and the patients who historically received the alternative treatment are equivalent. A historical control trial typically cannot ensure the equivalence of these groups. One reason, among others, is that patients who are selected for the current study (and who agree to “try out” the new treatment) may represent a special group; these patients, then, would not be comparable to the more typical patients used as historical controls (see Pocock, 1983; Gehan, 1984; Sacks et al., 1983).

Numerous other evaluation designs, termed “quasi-experiments” (see Cook and Campbell, 1979), include the one-group pretest-posttest design, the regression-discontinuity design, and many, many others. Despite differences in these designs (and in their rigor), they share with historical trials an inability to ensure the isolation of the treatment effect.

---

## Randomized Studies

The 1920s brought a significant breakthrough in controlled trials: the randomized design (R.A. Fisher, 1925, 1935<sup>8</sup>). Relying on a chance process (randomization) for assigning current patients to two alternative treatments ensured that the only source of differences between the two groups, at baseline, would be chance.<sup>9</sup> Thus, one could assess the likelihood that an observed difference in outcomes occurred because of the difference in treatments rather than chance. Given an objective outcome measure, one could be relatively sure of the validity of this assessment.

Strictly random procedures preclude the possibility that investigators would (even unconsciously) assign the healthier patients to receive the new treatment or usual care. In addition, there is a statistical expectation of equivalence in the two groups. Probability theory allows investigators to estimate the size of the average difference in outcomes that might occur by chance alone—provided that a random process was used to assign treatments to individual patients. Thus, only a difference in outcomes that is larger than would be expected on the basis of chance alone will be interpreted as a statistically significant indicator of a treatment effect.

---

<sup>8</sup>The development of the randomized design has also been linked to earlier work by Neyman (1923), given a newly translated version of Neyman’s 1923 theoretical work (see Speed, 1990; Neyman, 1990; Rubin, 1990b).

<sup>9</sup>Randomization of treatment assignment can be accomplished by flipping a coin, drawing lots, using a table of random numbers or using an appropriate computer program.



The first full-fledged “random clinical trials” in the medical area were conducted in the 1940s (Pocock, 1983, citing Medical Research Council, 1948 and 1950). Two well-known examples of randomized studies conducted since that time are the Salk vaccine placebo trial (Francis et al., 1955; Meier, 1972) and the long-term breast cancer chemotherapy trial (B. Fisher et al., 1975). Specifically:

- Rates of poliomyelitis among schoolchildren assigned the Salk vaccine were compared to rates among those receiving a “placebo.”<sup>10</sup> Assignment to vaccine or placebo was blinded, and the use of coded supplies in a predetermined order was the virtual equivalent of random assignment.<sup>11</sup>
- Ten-year survival of breast-cancer patients randomly assigned to a new treatment, consisting of surgery and adjuvant chemotherapy, was compared to outcomes of patients assigned to the usual treatment, consisting of surgery only.

Randomized clinical trials (or, more simply, randomized studies) have been deemed, statistically, the most suitable design for addressing the question of a medical treatment’s efficacy (see, for example, Mosteller et al., 1985). But such studies are not without weaknesses.

The weaknesses of randomized studies derive primarily from their typically high cost.<sup>12</sup> Because of their high cost, randomized studies have been used to formally evaluate relatively few medical interventions. Further:

- the randomized studies that have been conducted have typically focused on narrow questions (such as, what is the effect of 50 mg of a particular drug for patients in a limited age range; for example, 45-55 years of age); and

<sup>10</sup>A placebo is an inert substance or procedure given to subjects assigned to the control group of some experiments. The purpose is to prevent the patients themselves, the physicians judging patients’ outcomes, or both from being influenced by knowing which treatment each patient received (Kirsch and Weixel, 1988). Depending upon the treatment in question, placebos may or may not be feasible. The inability to use a placebo control (i.e., the inability to “blind” patients or physicians judging outcomes) increases the uncertainty associated with the results—to a marked degree for subjective outcomes, such as perceptions of pain, and to a lesser degree for objective outcomes, such as survival.

<sup>11</sup>In addition to the placebo trial, the Salk vaccine study included a nonrandomized study in which second-grade children were given the vaccine and their outcomes were compared to first- and third-graders who were not given the vaccine (see Meier, 1972).

<sup>12</sup>Other weaknesses do exist. For example, a randomized study is not suitable, first, from an ethical viewpoint, wherever there would be negative effects on some participants, and second, from a practical policy viewpoint, wherever there is a lengthy waiting time involved in determining future patient outcomes (Rubin, 1974).

- many randomized studies have been able to enroll only small numbers of patients.

These weaknesses have long been recognized (see, for example, Rubin, 1974).

Indeed, in the interests of achieving a more tightly controlled comparison, investigators who can afford to enroll only a small number have deliberately drawn study participants from a limited subset of patients. (If all patients are similar, then by definition, the comparison groups are composed of comparable patients.)

Elderly patients and others with comorbidities are often excluded from randomized studies for two reasons. First, investigators fear that adverse reactions to a new treatment may be greater among “weaker” patients. Second, investigators expect that a greater number of such patients will die during the study from other diseases and conditions.<sup>13</sup> In all cases where certain types of patients are excluded, there is a threat that results might not be generalizable to all patients.

The generalizability of results can vary greatly from one study to another. For example:

- Some randomized studies are conducted on patients who approximate a “microcosm” of the total patient population, thus allowing generalization from the observed results. One example is the placebo trial in the Salk vaccine study cited above—but, even here, there were some differences between study participants and those schoolchildren who did not participate.
- More often, however, randomized studies totally exclude certain patient groups. A well-known example is the MRFIT<sup>14</sup> study (Harrison and Morgan, 1986; Neaton, Grimm, and Cutler, 1987), which excluded women.<sup>15</sup>

---

<sup>13</sup>Indeed, an increasing likelihood of death for other reasons characterizes each older age group of patients. As Zelen and Gelman (1986) point out, when the key outcome of a randomized study is survival, the inclusion of older patients is associated with the dilution of results. For example, if a 40- to 44-year-old woman diagnosed with node-positive breast cancer dies within a 10-year period, the chances are 83 to 1 that the cancer was the cause of death; by contrast, for a similar woman aged 75-79 at diagnosis, the chances are only 1.3 to 1 that the cancer was the cause of death. Thus, much larger samples are required to achieve significant results for elderly patients.

<sup>14</sup>The Multiple Risk Factor Intervention Trial targeted males at risk of heart disease.

<sup>15</sup>When certain kinds of patients are excluded from a trial, it is often not known whether or not these patient groups would respond to the treatment in a similar fashion to those who were included.



- Other randomized studies include key patient groups, but underrepresent or overrepresent their relative numbers. For example, a randomized study may draw participants predominantly from younger age ranges, even though the patients that treatment is intended for are predominantly elderly.<sup>16</sup>
- In still other studies, hidden biases affect the patient selection process, and it may not be obvious—even to the investigators themselves—exactly which groups of patients are fully represented, which are underrepresented, and which (if any) are completely excluded.

Of course, wider recruitment may not always be an appropriate solution, especially for relatively small-sample studies. But this does not change the fact that the generalizability of results is a potentially important problem in each such study.

---

## Meta-analyses

The limited size and scope of many individual studies has been associated with inconclusive and, sometimes, seemingly inconsistent findings. This has led to the development of a new form of research that is still undergoing refinement. “Meta-analyses” (Glass, 1976; Glass, McGaw, and Smith, 1981; Hedges and Olkin, 1985; Wachter and Straf, 1990), or “quantitative overviews” as many medical researchers call them (Yusuf, Simon, and Ellenberg, 1987; Ellenberg, 1988; Peto, 1987), expand knowledge by statistically combining the results of multiple studies—often randomized studies—that all address essentially the same research question (Ellenberg, 1988).

As explained by T. Louis, Fineberg, and Mosteller (1985, pp. 1-2):

“A meta-analysis is to a primary research study as a primary research study is to its study subjects. The meta-analyst first decides on a research question (such as the relation between diet and hyperactivity), searches the literature for relevant primary studies, and then summarizes these studies with both descriptive and numeric presentations. Just as the primary study requires an admission rule for patients, so the meta-study requires an admission rule for primary studies (what sorts of primary studies should be combined?), a measurement rule (what endpoints are of primary interest and how should they be quantified?), and an analysis plan (what summaries, statistical comparisons, and estimates should be produced?).”

---

<sup>16</sup>For examples in the area of coronary artery disease, see Califf, Pryor, and Greenfield (1986).

Meta-analysis was first practiced by social scientists who used a standardized “effect size” to combine results from studies with different outcome measures (for example, different measures of self-esteem). Later, many medical researchers combined only those studies that had the same “endpoint” (for example, 5-year survival) or only those with random assignment or only those that met both criteria.<sup>17</sup>

By combining the results of many similar, small studies, using rigorous methods, the meta-analyst can approximate the results of a single large study at a tiny fraction of the cost of fielding a new large study. Thus, the most widely recognized advantage (or strength) of meta-analysis is the increase in sample size—and thus in statistical power—that a “pooled” estimate affords (Light, 1984). Within any one study, the numbers of patients in a subgroup may not be large enough for a stable estimate of the treatment’s effect in that particular subgroup,<sup>18</sup> but using meta-analysis to combine results observed for that subgroup in multiple studies allows greater stability.

Further, the meta-analyst can combine the results of several randomized studies, each of which covers different portions of the patient population. For example, he or she might combine results of a study that covers male patients (only) with another that covers females only. The combined results of such studies more closely approximate the total patient population.

Light (1984; Light and Pillemer, 1984) has pointed to another advantage of meta-analysis. The differences in various primary studies’ estimates of a treatment’s effect may have come about because of study-to-study variations in the research process. This strength of meta-analysis is illustrated by Lipsey’s (1992) analysis of a sample of 443 delinquency treatment studies. Using hierarchical multiple regression, Lipsey explored the impact of cross-study differences (for example, differing characteristics of the delinquents in each study) on the observed effect of the treatment. His objective was to differentiate the unique impact of each type of cross-study difference. (For example, the impact of differing characteristics of delinquents would be distinguished from the impact of

---

<sup>17</sup>In England, the “Oxford group” (Richard Peto and others) were among the leaders of this effort (see Ellenberg, 1988). A recent example of this approach is the two-volume set of meta-analyses assessing treatments in pregnancy and childbirth (I. Chalmers, Enkin, and Keirse, 1989). For a methodological review of meta-analyses of randomized studies, see Sacks et al. (1987).

<sup>18</sup>For example, within a single breast cancer study, the number of premenopausal women with three or more positive nodes may be too small to support a separate estimate of the benefits of adjuvant chemotherapy.



other cross-study differences, such as sample size, attrition, amount of treatment given.)

Building upon work by Light and by Lipsey, meta-analysts in the medical area could investigate, for example, whether studies evaluating a treatment on healthier patients show greater (or less) benefit than do studies conducted on more severely ill patients or those with comorbidities.

Rating the quality of each study that is a candidate for inclusion in a meta-analysis has been a focus of many working in this field (see T.C. Chalmers et al., 1981). Some meta-analyses have examined the relationship between the size of the observed treatment effect and the quality score of the study. Others have excluded studies deemed to be of lower quality or have, in effect, “written off” such studies (see Cordray, 1990b). Rubin’s (1990a) approach, described below, includes all studies.

The most obvious limitation of meta-analyses is that the results do not extend beyond the set of studies that have already been conducted. A meta-analysis of randomized studies cannot cover elderly patients if no (or virtually no) elderly patients were included in these studies. And the stricter the inclusion criteria, the more limited the pool of qualifying studies and available results (see Cordray, 1990a, 1990b). As previously noted, many meta-analysts assessing medical treatment effects have excluded all nonrandomized studies; however, to capture more information, some have included more diverse studies. For example, American social scientists working in the medical area (Wortman and Yeaton, 1983) included results of both randomized studies and nonrandomized “quasi-experiments”; results were stratified by type of study design.<sup>19</sup>

Critics of meta-analysis fear that the weaknesses and biases of existing studies may be compounded when they are quantitatively combined. For example, individual studies on a topic—far from being independent—may have been conducted by a single researcher and his (former) students. A convergence of results across such studies gives the appearance that a certain finding has been independently confirmed; in reality, such convergence may be merely an artifact, arising from masked correlations

---

<sup>19</sup>Meta-analyses focusing not on treatments but on risk factors for various diseases have combined data from epidemiologic studies, which are not randomized. One example of such a study is the meta-analysis of studies of the relationship of blood pressure to stroke and coronary heart disease (MacMahon et al., 1990). Another is the meta-analysis of studies of the relationship between alcohol consumption and breast cancer (Longnecker et al., 1988).

between results and the affiliation of the researchers involved. (Such problems would be mitigated by combining diverse types of studies that are more likely to have been conducted by independent investigators and to have complementary strengths and weaknesses.)

As pointed out by Rubin (1990a), much meta-analysis has been oriented toward summarizing existing studies—rather than toward using existing studies to discover scientific reality. The distinction is an important one: Typically, meta-analysts seek to include all studies that have been conducted on a topic (or all that meet pre-set inclusion criteria), thereby representing the work that has been done in the field. By contrast, few seek to ensure that all relevant patient population groups are adequately represented; yet such representation is an equally—or more—important sampling issue (Hedges, 1990; Light and Pillemer, 1984).

Most recently, alternative or expanded ways of synthesizing study results have been suggested by Eddy (Eddy, 1989; Eddy, Hasselblad, and Shacter, 1989; Eddy et al., 1988) and by Rubin (1990a)—and have also arisen in the work or comments of other researchers in the medical field.

Specifically, Eddy has constructed models interrelating “pieces of evidence”; that is, results from diverse studies that may address different, but related, research questions. Results of studies estimating numbers of women receiving mammography might be combined with studies of its effectiveness in diagnosing breast cancer and with studies of the impact of early diagnosis on patient survival. Eddy’s work focuses on ways of adjusting and combining the results of diverse studies that address these questions, given knowledge of certain biases; it does not focus on ways of determining the nature and degree of bias in each study.

Rubin (1990a) has proposed that rather than exclude all but the highest quality randomized studies, the meta-analyst should use all existing studies to project results for an “ideal study.” Rubin’s reasoning is that all studies are imperfect and that extrapolation to the ideal study would, as appropriate, give more weight to higher quality studies. Presumably, the diverse set of studies to be included (randomized and nonrandomized) would allow the projection to cover the full population of patients. Glass (1991) endorses Rubin’s view.

Mosteller (1990b, p. 189, citing remarks of Singer) has also suggested the possibility of synthesizing information from diverse sources, although he



notes that “executing such a maneuver may be challenging, and convincing others of its validity even harder.”

In a somewhat similar vein, a recent study by the Oxford group (Collins et al., 1990) included (1) a traditional meta-analysis of randomized studies of the impact that blood pressure drugs have on stroke and coronary heart disease, and (2) a separate presentation of data from epidemiologic studies of the natural or usual relationship of blood pressure to stroke and coronary heart disease. The epidemiologic data set the “context” for randomized studies’ estimates of drug effects.

Other meta-analysts attempting to compare the “relative merits of allogeneic bone-marrow transplantation . . . and conventional chemotherapy” (Begg and Pilote, 1991, p. 899; see also, Begg, McGlave, and Pilote, 1989) found no published randomized studies, a few small, nonrandom comparative studies, and a number of larger uncontrolled studies. They therefore explored ways of combining the existing nonrandom information derived from a variety of less than optimal study designs.

---

## Data Base Analyses

With the advent of the computer age, another type of study has appeared: the data base analysis.<sup>20</sup> Computerized data bases routinely maintain records for thousands of patients. In many data bases, the record for each patient includes information on his or her diagnosis, treatment, and outcome (see Pryor et al., 1985; Connell, Diehr, and Hart, 1987; Tierney and McDonald, 1991). In recent years, such data bases have often been made available for public use, with information identifying specific patients stripped away. Recently, analysts interested in medical effectiveness have begun to use these data bases (see Ellwood, 1988; Roper et al., 1988; McDonald and Hui, 1991).<sup>21</sup>

Nonrandomized studies have long been recognized as generally providing superior coverage (see Rubin, 1974). Many observational data bases capture the treatments and outcomes that occur in the normal course of medical practice—and a fuller range of patients than any selective study

---

<sup>20</sup>Before computer technology for storing and retrieving patient information, certain hospitals and clinics archived patient records. Notably, the Mayo Clinic kept records on individual patients from the beginning of this century (Pryor et al., 1985, p. 631, citing Kurland and Molgaard, 1981, and Bill et al., 1978).

<sup>21</sup>This form of research has often been referred to as “outcomes” research because it explores medical effectiveness by comparing the outcomes of patients receiving a new treatment to outcomes for other patients receiving usual care.

could approximate. For example, the Medicare data base includes the vast majority of all elderly patients in the nation.<sup>22</sup> The SEER Data Base includes all cancer patients entering all hospitals in several large geographic areas.<sup>23</sup> The Duke Databank for Cardiovascular Disease includes all patients with chest pain at Duke University Medical Center (Pryor et al., 1985). It contains a wealth of data on each patient's demographic profile, diagnosis, treatment, and eventual outcome. By contrast, data bases comprised of patients participating in a group of selective studies (see, for example, Moon et al., 1987) would clearly not qualify as approximating the full range of patients.<sup>24</sup>

A recent study by Krakauer (1986) provides an example of a data base analysis that evaluates treatment effectiveness. This study of end-stage renal disease patients who had received transplants compared two alternative immunosuppressive regimens: cyclosporin (an expensive drug) and conventional therapy. Using a data base jointly developed by the Health Care Financing Administration (HCFA) and the Social Security Administration, Krakauer compared outcomes (graft retention and costs) for the two groups of patients; the results showed that cyclosporin reduced graft failure—and did so without incurring additional cost (because the need for dialysis was also reduced). The relative risk of graft failure was used to estimate the treatment effect. In estimating the relative risk, adjustments were made equating the comparison groups on age, race, various prognostic factors, and comorbidities (for example, diabetic nephrosclerosis).

Data base analyses (or outcomes research) have a number of attractive characteristics. An obvious advantage is low cost, because the data have already been collected. Another advantage is that many data bases cover the full range of patients receiving the treatment in medical practice. This is important, given the limited coverage in randomized studies and even meta-analyses of randomized studies.<sup>25</sup>

---

<sup>22</sup>The Medicare data base is maintained by the Health Care Financing Administration (HCFA), which is part of the U. S. Department of Health and Human Services. It is sometimes referred to as the HCFA data base.

<sup>23</sup>The Surveillance, Epidemiology, and End Results (SEER) Data Base is maintained by the National Cancer Institute.

<sup>24</sup>Of course, a researcher should assess the coverage that each data base provides for the specific patient population in question. For example, hospital data bases may be far from complete for conditions and diseases where substantial numbers of patients are not hospitalized.

<sup>25</sup>Still other advantages of data base analyses include (1) their timeliness (since the outcomes of treatments have already occurred and there is no need to wait, e.g., 10 years, to assess long-term survival); and (2) their freedom from the ethical concerns that manipulation of treatments in randomized studies sometimes involves (see Rubin, 1974).



Despite these advantages, the outcomes approach suffers from several potential weaknesses. These include limited patient descriptors, potential recording and transcription errors, and missing data, to name a few.<sup>26</sup> Given our focus on estimating treatment effects, the chief weakness of data base analyses—and indeed all observational studies—is undoubtedly the likelihood that the patient groups being compared were not comparable at baseline. If so, then differences in outcomes for patients who received different treatments may be coincidental to and not caused by the difference in treatments.

This problem is referred to in this report primarily by the term “comparison bias.” Other terms, used in this report to refer to specific aspects of this problem, include: “treatment assignment bias” and “imbalanced comparison groups.”<sup>27</sup> The potential for comparison bias can be appreciated when one realizes that treatment assignment is often biased: the healthier patient, who has a relatively mild case of a disease and a high likelihood of a favorable outcome, may be considered a “good candidate” for a certain treatment and is therefore likely to receive it. By contrast, the patient with a more severe case and a worse prognosis may be assigned to an alternative treatment.<sup>28</sup> The upshot is, oftentimes, imbalanced comparison groups in data base analyses.

To adjust for initially imbalanced comparison groups, analysts have turned to a variety of statistical methods.<sup>29</sup> But unfortunately, no after-the-fact adjustment can be counted on to approximate the assurance of equivalence that randomization provides.

## Implications of the Review of Designs

The foregoing review of study designs indicates, first, that two sets of factors contribute to the results of any study. These are (1) the real effectiveness of the treatment, and (2) the “filter” imposed by the design of

<sup>26</sup>A particularly vexing problem for many data base analysts is the lack of follow-up as individual patients move in and out of the “net” or “capture area” of a particular data base. However, some data bases are not affected by this problem (Medicare), and others solve it through telephone follow-up (Duke).

<sup>27</sup>Still other terms for this problem, which are not used in this report because of the potential for confusion, include “selection bias” (Byar, 1980) and “nongignorable treatment assignment” (Rubin, 1978). Selection bias can refer to either (1) selection of a nonrepresentative sample for a study or (2) biased selection of individuals who will receive a particular treatment—or both. “Nongignorable treatment assignment” means that the process by which treatments were assigned cannot be ignored when making a comparison of outcomes.

<sup>28</sup>For example, the stage of a cancer patient’s disease often determines which therapy is advised.

<sup>29</sup>This is sometimes referred to as “controlling for confounding factors.”

a less-than-perfect study. For example, a given study result (such as: “there is a 50-percent reduction in tumor size for the treatment group”) has been produced not only by the reality of the effect of the treatment on tumor size, but also by such idiosyncracies as who participated in the study, who was assigned to which treatment, how tumor size was measured and recorded, the type of analysis that resulted in the 50-percent figure, and possibly other factors.

Thus, to the degree that a study’s weaknesses and biases are not understood, the meaning of that study’s results is uncertain. Each study’s strengths and weaknesses should be assessed and taken into account when interpreting its results, in order to reduce or minimize the potential for misinterpretation. Further, no single study—indeed no single type of study design—is likely to reveal the complete effectiveness domain clearly.<sup>30</sup>

Secondly, the review of designs indicates that combining research results through meta-analysis has a number of advantages. Meta-analysis is less expensive than studies that require new data collection, and it often allows one to reach broader conclusions than are possible in a single study. As typically practiced thus far, meta-analysis has limited ability to answer our question on the effectiveness of a treatment across all patients. However, efforts to synthesize study results are still undergoing refinement and change.

Thirdly, the review of study designs indicates that the strengths and weaknesses of different designs are often complementary. Two very different study designs—randomized studies and data base analyses—have complementary strengths and weaknesses.<sup>31</sup> By definition, the primary strength of the randomized study is controlled comparison, whereas the chief weakness of the data base consists of uncontrolled and potentially imbalanced comparison groups (McDonald and Hui, 1991). And given the effectiveness question under study in this report, the primary

---

<sup>30</sup>Recently, well-known analysts (Richard Peto of Oxford and Paul Meier of the University of Chicago, among others) have called for exceedingly large controlled studies. Indeed, some have informally stated a belief that every patient in the United States should be enrolled in a randomized trial. The costs of such large enrollments might be offset to some degree by collecting information on fewer variables, but this would limit possibilities for estimating differential effects of the treatment in different patient subgroups. (It would also limit estimation of differential effects for different implementations of the treatment and for different outcome criteria.) And even if measurement of fewer variables were deemed appropriate, it seems unlikely that vast undertakings would be feasible for most medical interventions—or that patients would agree to accept random assignment to treatments.

<sup>31</sup>Other researchers—notably, Califf, Pryor, and Greenfield (1986)—have also viewed randomized studies and data base analyses as complementary designs.



strength of a data base analysis lies in its ability to capture all aspects of actual medical practice—in particular, the full range of patients for whom a treatment is intended or actually used; by contrast, the key weakness of randomized studies is their potential lack of generalizability (see Califf, Pryor, and Greenfield, 1986).

In other words, when well conducted, randomized studies should provide the valid comparisons needed in scientific evaluations; however, if such studies provide only limited coverage of the effectiveness domain, their results may not be sufficiently general to serve as a basis for setting national guidelines for medical practice. In such situations, it may be that data base analyses can be used to complement randomized study results.

The primary strengths and weaknesses of randomized studies and data base analyses are summarized in table 1.1.

Table 1.1: Complementarity of Strengths and Weaknesses of Two Study Designs

Study design	Primary strength	Primary weakness
Randomized studies	Controlled comparison; internal validity	Potential lack of generalizability; external validity at risk
Data base analyses	Coverage of medical practice (full patient population, full range of treatment implementations); external validity	Uncontrolled comparison; internal validity at risk

Drawing upon the results of studies with complementary strengths and weaknesses logically has the potential to yield more—and better—information than could be provided either by a single study or by a group of studies that share essentially the same strengths and weaknesses. There is a potential benefit to strategically combining the results from diverse studies that have different, complementary strengths and weaknesses, but that are nevertheless alike in their goal of estimating the same treatment’s effect.

Of course, one cannot simply average studies together, hoping that the different biases will counteract (or “cancel out”) one another. Nor can one assume that the different biases are all equally severe. Thus, the potential of a synthesis of complementary studies can be realized only if the diverse strengths of the different studies can be captured, while recognizing each study’s specific weaknesses and minimizing their impact on results. This is

the goal of the strategy presented here, which we term “cross design synthesis.”

## Objective, Approach, and Methodology

Our overall objective is to develop a strategy for cross design synthesis: specifically, a methodology for combining results from diverse, complementary studies that have evaluated a given treatment's effect. The long-term goal of our work is to improve knowledge about the effectiveness of medical interventions and thus to enhance the development of valid practice guidelines. Specifically, by increasing the validity, the credibility, and the efficient dissemination of such information, the potential for successful medical practice guidelines in the United States is enhanced.

Our approach is anchored in meta-analytic principles and techniques. However, a cross design synthesis is quite distinct from a traditional meta-analysis. Foremost among the differences is the fact that a cross design synthesis combines results from study designs that have complementary strengths and weaknesses, and it is specifically aimed at reaping the benefits of the studies' diverse strengths, while minimizing the bias associated with each study. The spirit of such an effort is well described by Shadish, Cook, and Houts (1986, p. 43) in their discussion of quasi-experimentation (that is, imperfect designs) from a “critical multiplist perspective”:

“From a critical multiplist perspective, ... [dealing with imperfect designs] ... resembles chess in several ways. Each chess piece has different strengths and weaknesses with respect to mobility and direction. Similarly, ... no two kinds of design or analysis are the same; each is associated with a unique set of biases. In chess, no single piece is capable of winning the game by itself. Rather, the best chance of winning occurs when all the pieces work together, using the strengths of each to protect against the weaknesses of the others.”

As already noted, the challenge of cross design synthesis lies in the fact that one cannot expect the various design weaknesses (and consequent biases in study results) to cancel out one another. This makes any attempt to combine the results of different, complementary types of studies difficult: The different weaknesses associated with the different designs must be assessed according to a specific strategy and then taken into account when combining the results.

In light of this, strategies for cross design synthesis must be based on the full set of methods that have been used to assess, adjust, and combine



studies. A cross design investigator must act within each of these methodological areas. However, the most important for cross design synthesis may be assessment; that is, discriminating between study results that speak to the true effectiveness of an intervention and results that merely reflect consequences of how studies were conducted. What is learned in the assessment of existing studies determines whether a cross design synthesis is needed, possible, and advisable. Further, the results of assessment will guide strategies for adjusting and combining studies. Thus, although this report reviews methods for assessing, adjusting, and combining study results, its greatest emphasis is placed on methods for assessing study weaknesses.

Given that a cross design synthesis will include studies with very different kinds of weaknesses and potential biases, different assessment methods are appropriate for the different types of studies to be included in the synthesis. This, in turn, argues for limiting the numbers of different types of studies to be included in a synthesis and for focusing on only the primary weaknesses of those studies (at least in this first report).<sup>32</sup> Therefore, in this initial presentation of the cross design synthesis, we limit our efforts to two types of study designs—randomized studies and data base analyses—and we focus on the primary weakness associated with each of these designs.<sup>33</sup>

Although our main reason for selecting these two designs is their complementarity, a second reason is their current prominence in the field. Doubtless, the randomized design is considered the “gold standard” for evaluating medical interventions. And the recent proliferation of data base analyses combined with funding opportunities for outcomes research make it likely that such analyses will increase in importance over time. Other types of studies (for example, historical control trials and other nonrandomized designs) might also have been considered for a cross design synthesis. These could well be the focus of future work in this area.

In sum, the strategy of cross design synthesis presented here is limited to methods for assessing, adjusting, and combining treatment effects from

---

<sup>32</sup>The primary weaknesses are defined in terms of the research question being addressed.

<sup>33</sup>Others have also recognized that strategies involving these two study designs provide certain benefits. Notably, Krakauer and Bailey (1991) have put forward a prospective strategy for sequentially conducting and planning research. In the Krakauer-Bailey model, analyses of data bases would be conducted first, and then, where justified by the results of these analyses, randomized studies would be planned. Moffitt (1991) has suggested that prospective approaches could include not only the sequential plan but also the simultaneous fielding of randomized studies and data base analyses. The simultaneous studies could be used to test hypotheses about the effect of study design on results.

studies of two major designs: specifically, randomized studies and data base analyses. The specific objectives of this report are further limited to resolving the chief weakness associated with each of these designs; that is:

- in randomized studies, the lack of generalizability (which is an issue of external validity); and
- in data base analyses, imbalanced comparison groups (which is an issue of internal validity).

Once we decided upon our specific objectives and our general approach, we engaged in a sequential process that can best be characterized as informed trial and error. We began by reviewing the relevant methodological literature, including:

- methods for assessing generalizability of randomized studies' results,
- methods for assessing imbalanced comparison groups in data base analyses (or other nonrandomized studies), and
- methods for adjusting and combining research results.

We limited our review of methods for assessing randomized studies to methods for assessing generalizability; similarly, our review of methods for assessing data base analyses is limited to methods for assessing comparison bias. We therefore assumed that the studies were otherwise well conducted. In an actual application of cross design synthesis (as in traditional meta-analysis), the investigator would assess studies on numerous criteria (see, for example, Himel et al., 1986). The in-depth assessments of key weaknesses described in this report are intended as an addition to the more usual assessments; those assessments typically cover a comprehensive set of potential biases, but subject each to relatively superficial scrutiny.

Studies that were not well conducted might be eliminated from a synthesis on the basis of relatively superficial assessments. But the point of the in-depth assessments described here is that randomized studies' results will be included in the cross design synthesis even though it is known in advance that they are probably not generalizable. The reason is that they have an important strength in another area. In order to minimize the weakness (lack of generalizability), the investigator must have an in-depth understanding of its nature and extent in each randomized study. The situation is similar for data base analyses and the potential problem of imbalanced comparison groups.



In reviewing methods for assessing, adjusting, and combining studies, our purpose was to identify as full a range of analytic procedures as possible. To this end, we conducted computerized literature searches, hand-checked key medical journals (especially letters to the editor, which we deemed a possible source of informal methods of assessing published studies), and asked numerous consultants to suggest “leads.”<sup>34</sup> We drew material from the literature of meta-analysis, evaluation research, general social science research methods, applied statistics, and substantive medical research.<sup>35</sup>

We then organized our findings according to frameworks that would help us identify gaps in available methods. Where we initially found gaps, we renewed efforts to find previously published methods that might fill these gaps.

Finally, we drafted a set of tasks and steps for conducting a cross design synthesis based on existing methods of assessing, adjusting, and combining studies. We then had these steps extensively reviewed by individuals expert in the major types of effectiveness studies.

---

## Organization of the Report

The three subsequent chapters in this report parallel the tasks one would follow in performing a cross design synthesis. Chapters 2 and 3 present methods for assessment, based on approaches reported in the literature. Specifically, chapter 2 reviews methods for the task of assessing the generalizability of randomized studies’ results. Chapter 3 reviews methods for the task of assessing imbalanced comparison groups and resulting bias in data base analyses. Chapters 2 and 3 provide the basis for the subsequent tasks: adjusting the results of existing studies, as appropriate, and strategically combining these results, while taking account of differences across studies. Specific methods for adjusting and combining results are discussed in chapter 4.

---

<sup>34</sup>See Appendix I: List of Experts.

<sup>35</sup>We did not cover other fields such as astronomy, physics, or the philosophy of science.

# Methods for Assessing Randomized Studies' Generalizability to the Patient Population (Task 1)

---

Randomized clinical trials (or more simply, randomized studies) are designed to achieve a valid comparison of alternative treatments, that is, internal validity.<sup>1</sup> By contrast, lack of generalizability (external validity) is not specifically addressed by the randomized study design and, as explained in chapter 1, constitutes the chief potential weakness associated with such studies' results. The treatment effect observed for the perhaps select subset of patients participating in a randomized study may or may not be generalizable to all relevant patients. Yet knowing the effects of a treatment in various patient subgroups and the degree to which those effects will be observed in the full population is essential for setting medical practice guidelines.

Therefore, to minimize the risk of overestimating or underestimating a treatment's average effect in the patient population, results of existing randomized studies should be subjected to an in-depth assessment of their generalizability. In conducting this assessment, the specific purposes of the cross design investigator are to:

- determine whether there is a need for a cross design synthesis (that is, whether there is a need to draw upon studies of other designs to expand the generalizability of existing randomized studies' results);
- provide a basis for adjusting each randomized study's results, where necessary, so as to enhance their generalizability; and
- inform judgments about persistent limitations of each study (including judgments about the level of uncertainty associated with the each study's results), which will guide choices among methodological options when—later—synthesizing results of diverse studies.

Thus, the first major task facing the cross design investigator is to assess the generalizability of randomized studies' results. Previous work in assessing the generalizability of existing studies' results falls into two major categories. They are:

- Assessment of the methodological process by which patients were selected into a randomized study, based on logic and subjective reviewer judgments about whether the process was aimed at achieving a representative patient pool.

---

<sup>1</sup>The discussion in this chapter assumes the randomized studies in question were well conducted, thus reaping the benefits of internal validity that are possible with randomized design. Of course, the extent to which any one specific randomized study actually achieves internal validity depends on a number of factors (see, e.g., Pocock, 1983). The Chalmers' quality rating scale for randomized studies includes several internal validity items (T.C. Chalmers et al., 1981).



- Assessment of the result of the randomized study's patient selection process (that is, the representativeness of the achieved patient pool) based on empirical evidence of whether the patient pool includes the kinds of patients seen in medical practice.

## Judgmental Assessments of Patient Selection Methods

Reviewers' assessments of patient selection methods are necessarily subjective and are most useful when combined with the more objective and empirical assessments of representativeness discussed in a subsequent part of this chapter. Nevertheless, methods of patient selection are recognized as an important determinant of generalizability. They have been assessed in previous reviews of randomized studies. And, if the reviewers' assessments are properly conducted, they should indicate whether the selection of patients for the study was aimed at attaining a "microcosm" of the population of patients seen in medical practice.

We believe a properly conducted methodological assessment should cover all phases of the patient selection process. At least four phases have been identified:

- Phase 1: The target group is set (that is, patient selection criteria are defined).
- Phase 2: The patient recruitment mode is chosen and activated.
- Phase 3: Some patients who met the initial selection criteria are subsequently rejected, on a case-by-case basis, by the study investigator.
- Phase 4: Eligible patients prove willing (or unwilling) to participate when selected. In some instances, they avoid selection.

The following sections discuss each of the four phases of the patient selection process and ways of improving assessments of this methodological process.

### Phase 1: Target Group (Patient Selection Criteria)

The potential for nonrepresentativeness resulting from a limited target group has been emphasized by McPeck (1987). Similarly, Merigan (1990) and Byar et al. (1990) have recently called for wider target groups—specifically, in trials of treatments for acquired immunodeficiency syndrome (AIDS). As these methodologists recognize, target groups have often been limited intentionally, in an effort to increase the internal validity of the trial's results. But the upshot has been randomized studies conducted on nonrepresentative groups of patients.

One frequent mode of limiting the target group has been to exclude elderly patients—even when they constitute a substantial portion of the relevant patient population.<sup>2</sup> For example, elderly breast cancer patients have been excluded from major randomized studies, although 43 percent of all breast cancer patients are over 65 at the time of diagnosis (Lichtman and Budman, 1989, citing Yancik, Ries and Yates, 1989).<sup>3</sup> Another common mode of limiting the target group is to exclude women patients, as has occurred in many randomized studies in the area of heart disease.<sup>4</sup>

The Chalmers scale for rating the quality of a randomized trial includes an item on patient selection criteria (which define the target group). Specifically, this item requires the investigator to judge the adequacy—not of the selection criteria themselves—but of the description of those criteria that is included in the report of the randomized study (see T.C. Chalmers et al., 1981; Liberati, Himel, and T.C. Chalmers, 1986). Descriptions judged “adequate,” “fair,” or “inadequate” receive a prescribed number of points.

## Phase 2: Patient Recruitment Mode

The way that patients are recruited for a randomized study can seriously impair the generalizability of results. Pocock (1983, p. 36), for example, points to the source of patients—or recruitment mode—as a key issue in representativeness. He notes that:

“in the study of depressive illness if one recruits hospital in-patients one ends up with an atypical group. Such patients tend to be the more serious chronic cases whereas any new antidepressant drug is usually under investigation with an eye to the larger group of depressed patients....”

Thus, as Pocock stresses, patient recruitment must be representative; otherwise, the generalizability of the findings may be merely that of a “convenience sample.”

<sup>2</sup>The reasons why investigators exclude certain patient groups from randomized studies were discussed in chapter 1 of this report.

<sup>3</sup>The National Surgical Adjuvant Breast Project has generally excluded patients over age 70. This followed from a requirement for a life expectancy of 10 years or more, exclusive of a cancer diagnosis (see B. Fisher and Redman, 1989). The Ludwig Breast Cancer Study Group Trial excludes patients age 65 or older.

<sup>4</sup>See GAO (1990) testimony on problems of implementing the National Institutes of Health policy on including women in study populations.



---

### Phase 3: Rejection of Individual Patients Who Qualified

Remington (1989, p. I-67) has charged that “patient exclusion criteria represent a vast wasteland of clinical trial design.” He notes that, although investigators have been “relatively precise” in defining target groups and patient selection criteria, this has unfortunately not been the case for rejections of patients who had qualified under the initial criteria:

“... many [investigators] ... have been very crude in defining patient exclusion criteria ... in general ... [using only] a brief statement ... or such global phrases as ‘serious intercurrent disease.’”

Similarly, Liberati, Himel, and T.C. Chalmers (1986) report that fully two-thirds of the breast cancer treatment trials they assessed did not even mention having kept a log of patients rejected.

Exclusion of individual patients can have important cumulative effects on the representativeness of the entire patient pool. To cite one example, in reviewing the British Medical Research Council Trial, Remington (1989, p. I-67) points out that, of the 46,000 patients identified as eligible for this trial on the basis of measured blood pressure, more than half were excluded. Apparently, the excluded patients were those with worse-than-average prognoses since “ultimate mortality rates [were] much lower than expected in general populations with similar elevations of blood pressure.”

The Chalmers rating scale includes an item on each randomized study’s exclusions of patients; specifically, exclusions are judged on whether or not “a log [had been kept] of patients who had been seen by the investigators, but rejected before randomization as ineligible with listed reasons” (Liberati, Himel, and T.C. Chalmers, 1986, p. 945). Judgments of whether exclusions were adequately described—“Yes,” “Partial,” “No,” or “Unknown”—are associated with a prescribed number of points (see T.C. Chalmers et al., 1981, p. 46).

---

### Phase 4: Patient Willingness or Refusal to Participate

In certain illness or treatment areas, the reluctance of patients to participate in a randomized study (or of their physicians to refer them to the study) can seriously affect representativeness. Edlund, Craig, and Richardson (1985) call for primary investigators to provide more information on patient willingness to participate, noting that of 84 treatment studies reviewed, none had reported the number of refusers and nonrefusers.

Depending on the recruitment mode, patients (or primary physicians) may find it easy to avoid participation in randomized studies. In such instances, investigators cannot count individual refusals, but potential problems can be indicated by difficulties in achieving recruitment targets, as has occurred in randomized studies comparing lumpectomy with mastectomy (Taylor et al., 1984), or by low rates of participation in targeted clinics or hospitals, as has occurred in trials evaluating the intracranial/extracranial bypass (Barnett et al., 1987).

An example of the potential importance of patient refusals is Schooler's (1980, p. 30) description of a trial evaluating a drug used to prevent relapse in schizophrenia:

"The hypothesis to be tested was that guaranteeing receipt of medication would significantly decrease the number of patients who would relapse, and would also delay relapse for those who did ultimately relapse. Contrary to prediction, there were no differences in relapse rate between the two groups....[P]atients whose medication taking was controlled relapsed as early and as often as those who had to take oral medication daily.

"This study has been criticized on sampling grounds. It has been suggested that a significant treatment effect was not found because patients who refuse to enter a drug trial are the same patients who will be noncompliant with treatment, and that the restriction of the sample to those who consented to be studied also restricted the study to subjects who would take oral medication and therefore not show a treatment effect." (Emphasis added.)

---

## Improving Methodological Assessments

Representativeness of patients participating in a randomized study can be threatened by each of the four phases of the patient selection process—the target group definition, recruitment mode, rejection of individual patients, and patient willingness or unwillingness to participate. Alternatively, success in each phase can ensure representativeness. One example of a trial that comes close to ensuring representation in every phase would be the placebo trials that tested the Salk polio vaccine. These trials targeted all public schoolchildren in grades 1 through 3, recruited from all schools in communities at high risk of polio and apparently did not reject any of the children who appeared at the clinics. Further,



widespread cooperation was obtained from parents and children (Francis et al., 1955; Meier, 1972).<sup>5</sup>

All four phases should therefore be included in an assessment of the methods in the patient selection process. Current rating scale items (T.C. Chalmers et al., 1981) relating to generalizability are limited to only two phases of the patient selection process and, for these, to assessing whether or not each study recorded and described the relevant patient selection procedures as part of the report of results. The Chalmers' items, then, are aimed at assessing a study's documentation of procedures, rather than assessing the procedures themselves. In particular, the items do not assess the degree of representativeness or nonrepresentativeness that is implied by the study's procedures. One way to do that would be to derive separate, subjective rating scores for each of the four phases of patient selection discussed in the previous section. These might include ratings for both whether or not primary investigators had provided an adequate report of each phase and, where adequate reports were provided, the implied inclusiveness or selectivity of the procedures followed in each phase.

When applied to multiple existing randomized studies on a particular topic, descriptions of all phases of the patient selection process should reveal similarities and differences across individual randomized studies in the set. Similarities and differences in the portions of the patient population covered by different studies indicate where they overlap or where they are complementary.<sup>6</sup> In this way, a reviewer can determine whether the various randomized studies were nearly identical or extremely diverse, and perhaps complementary, in their selection of the patient pool.

<sup>5</sup>Over 60 percent of all children in grades 1 through 3 of the targeted schools provided signed parental permission slips to participate, and the vast majority of these actually received the full series of injections—with the remainder being classified as absent at the first or subsequent clinics or as withdrawals (see Francis et al., 1955, table 1a, p. 2). In other words, 100 percent of the relevant population was defined as the target group and 50 to 60 percent of the total actually participated. In considering this, one must recognize that in many other trials, participation is much, much lower than 50 percent. For example, a target group that excludes all women and elderly patients may represent only about 25 percent of the relevant patient population. Furthermore, perhaps only one-third of the target group will actually participate—as might be the case if patients (or primary physicians) have a strong preference for one or the other treatment and do not wish to be randomized or if investigators eliminate large numbers of eligible patients as “poor candidates.” This would yield a combined figure of only  $(0.33)(0.25) = 0.083$  or only about 8% of the relevant population.

<sup>6</sup>Coverage provided by two studies would be complementary if, for example, one covered women patients only, whereas another covered men patients only.

Judgmental assessments of the patient selection process, taken by themselves, are weaker than when combined with the empirical assessments described in the following sections. Empirical assessments are designed specifically to provide information about the consequences of patient selection procedures in randomized studies; that is, the representativeness of the achieved patient pool and the impact of any nonrepresentativeness on the generalizability of the observed treatment effect. We believe that empirical assessments of representativeness should be made unless the patient pool has been randomly selected from the reference population or approximates the total patient population.

---

## Empirical Assessments of the Representativeness of the Achieved Patient Pool

The representativeness of patients participating in a randomized study refers to the extent to which the achieved patient pool reflects the full population of patients for whom the treatment is intended. In assessing representativeness, it is unrealistic to expect (or to test for) achievement of a perfect miniature of the population (Kruskal and Mosteller, 1981); instead, one must ask whether the patient pool is representative with respect to a specific criterion. Because our goal is to estimate treatment effectiveness, the appropriate criterion is whether persons participating in the study are representative of the different ways that patients are affected by the treatment in question. Thus, the key question is: Does the treatment effect observed for participating patients represent the average treatment effect that would pertain for all patients in the relevant population?

Representativeness of the treatment effect cannot be measured directly; thus, existing assessments have necessarily proceeded via two-part approaches. The major two-part approach consists of (1) directly comparing the baseline characteristics (for example, patient sex and age) of participants in a randomized study to those of patients seen in medical practice, and (2) looking for evidence of whether each underrepresented and overrepresented subgroup (for example, elderly female patients) experienced a different treatment effect than other patients.

Another, less well-known approach also proceeds in two parts. As explained below, levels of patient outcomes (not effects) in each randomized study are compared to levels of outcomes for patients who received the same treatment in medical practice. Then, the reviewer looks for evidence that different levels of outcomes are linked to differential effects.



## Assessing Representativeness Using Patient Baseline Characteristics

Assessing representativeness via patient baseline characteristics is a two-part approach: Part 1 involves comparing baseline characteristics. Part 2 involves checking whether differences on these characteristics are linked to differential treatment effects. Part 1 and part 2 are discussed separately below.

### Part 1: Comparing Patient Baseline Characteristics

The representativeness of study subjects has often been assessed by comparing baseline characteristics of persons participating in a study with those of persons not participating. For example, a Swedish health intervention study (Wilhelmsen et al., 1976) used records of the local Temperance Board to compare the prevalence of alcohol problems among those who participated and among those who had refused or failed to appear for the study.<sup>7</sup> In a similar fashion, Steinhorn et al. (1983) compared the demographic characteristics of colon cancer patients in Comprehensive Cancer Centers with those of colon cancer patients in the SEER registry. Recently, Moon (1989) advocated comparing patients in a randomized study to those in a population-based registry (data base) in order to assess generalizability and improve the interpretation of results.

Comparison of baseline characteristics provides a mapping of the kinds of patients that are—and are not—covered by existing studies. Such a mapping might show, for example, that elderly patients were completely excluded from randomized studies of a certain treatment even if that age group had not been explicitly ruled out by the initial definition of the target group. Such a mapping would also indicate whether excluded groups constitute a majority or a small minority of all relevant patients in the population.

Any of several existing data bases could provide information to be used as a yardstick against which baseline characteristics of patients in randomized studies could be compared. A wide variety of data bases, such as SEER, are reviewed by Pryor et al. (1985), Mosteller et al. (1985), and Tierney and McDonald (1991).<sup>8</sup>

---

<sup>7</sup>They found an alcohol problem rate roughly three times higher among nonparticipants than among participants.

<sup>8</sup>Interestingly, for present purposes (i.e., a standard for judging coverage of the full patient population), a data base of patients and their characteristics need not include treatment information. For example, AIDS patients in a drug trial could be compared to AIDS cases in the data base maintained by the Centers for Disease Control. The fact that the CDC data base does not include treatment information is irrelevant to its use as a yardstick for assessing the representativeness of participants in a randomized study on measured baseline characteristics (such as sex). Of course, cases reported to CDC would ideally be adjusted for artifacts of the surveillance system (see GAO, 1989b).

In general, comparisons of patient baseline characteristics should cover all potentially relevant characteristics for which data are available. Such characteristics may be ranked as follows:

- The most important patient baseline characteristics are those that are known to be linked to differential effects of the treatment in question. Often, however, there is limited information regarding which patient characteristics predict above-average versus below-average responses to a new treatment.
- The next most important patient characteristics are prognostic factors; for example, indicators of severity of illness. Patient prognostic factors (such as the size of a cancer patient's tumor or the stage of a disease) have been linked to differential effects of specific treatments.<sup>9</sup>
- Finally, numerous demographic and other patient characteristics are also potentially related to the effect that a treatment will have. For example, studies of chemotherapy for breast cancer show that younger (that is, premenopausal) patients have a better response to chemotherapy than older (postmenopausal) patients (see B. Fisher et al., 1975).<sup>10</sup>

A reviewer rarely has access to full distributions of baseline characteristics; however, he or she can sometimes extend the foregoing comparisons of patient characteristics beyond measures of central tendency (such as the mean) by using reported measures of dispersion (such as the standard deviation). For example, the standard deviation of patients' ages may be much smaller in a randomized study than in a data base—even if both groups of patients have the same average age. Such a pattern would suggest that the randomized study may have excluded many elderly and youthful patients, choosing a more homogeneous “middle-aged” group.

Where possible, comparing full distributions is the best approach, since this identifies the degree to which patient groups are underrepresented or overrepresented. It also distinguishes those special instances where virtually no representatives of a patient group are included. The latter constitutes a more serious problem than instances where the

---

<sup>9</sup>For example, stage I breast cancer patients generally have better outcomes than stage II patients. This is particularly true for “low-risk” stage I patients, who, for example, have very small tumors. The stage of a patient's disease is related to the effect that a given treatment is likely to have on that patient. Specifically, chemotherapy combats systemic disease; therefore, stage II patients (who have relatively early systemic disease) are more likely to respond to chemotherapy than low-risk stage I patients (who generally do not have systemic disease).

<sup>10</sup>Another example consists of those patients with lower socioeconomic status or those without social supports, who might not respond to certain at-home treatments as well as other patients who have more resources to draw upon in their homes.



nonrepresentativeness is simply a question of relative numbers. In addition to comparing single-characteristic distributions, joint distributions of multiple patient characteristics (for example, age by stage of disease) could be compared (randomized study versus data base). Such a comparison might reveal, for example, that a randomized study includes no patients who are both old and severely ill, even though some old and some severely ill patients were included. However, reviewers are limited by available information.

If several randomized studies are being reviewed, the combined coverage of the set of randomized studies should be considered, particularly if complementary portions of the patient population are covered by different randomized studies. The reason is that even if no single randomized study covers the entire patient population, the combined set of existing randomized studies may come close to doing so. For example, if one randomized study covers male patients and a second randomized study covers female patients, together the two randomized studies cover both. This situation differs greatly from that in which all studies fail to cover the same patient group.

## Part 2: Checking Linkage of Baseline Characteristics to Treatment Effects

Once the reviewer has assessed the representativeness of patient baseline characteristics, he or she addresses the next logical question: whether any nonrepresentativeness that was detected for these characteristics actually translates into nonrepresentativeness with respect to the treatment effect. The danger is quite simply that a treatment may be more or less effective for the kinds of patients participating in randomized studies than for other kinds of patients.

As previously noted, patient nonrepresentativeness on such measurable variables as sex, age, or known prognostic factors may or may not translate into a lack of generalizability for the estimated treatment effect. For example, suppose that a randomized study's patient pool does not represent all patient ages or that it does not represent all stages of the disease being treated. If the treatment has different effects on patients in different age groups or on those with different stages of the disease, then the results of this study will not be generalizable to the full population of patients.<sup>11</sup> But if the treatment in question is equally effective for patients

---

<sup>11</sup>This would be the case if, for example, younger patients responded well and received a great deal of benefit, but older patients were helped little or not at all.

of all ages and for all stages of the disease, there is no reason to believe that the results of this randomized study are not generalizable.<sup>12</sup>

We therefore reviewed the literature on assessing linkage of the overrepresentation (or the underrepresentation) of a patient subgroup to a difference in the treatment's effect. We found two primary approaches to assessing linkage:

- One was to review subgroup analyses in published randomized studies to find comparisons of treatment effects across the specific subgroups that were found to be overrepresented (or underrepresented) in a randomized study.
- The other was to conduct meta-analyses, comparing treatment effects across randomized studies that overrepresent and underrepresent different subgroups of the patient population.<sup>13</sup>

Conducting a subgroup review to check linkage to differential treatment effects. A long-standing method of assessing linkage between known nonrepresentativeness (for example, on age, sex, or race) and differential treatment effects is to review existing randomized studies for reports of treatment effects within the relevant subgroups. The review of existing randomized studies amounts to a search for evidence that the treatment in question is—or is not—more (or less) effective for the overrepresented or underrepresented patient subgroups (for example, elderly, blacks, women). A subgroup review may help in specifying the kinds of patients for whom the treatment works best, least, and so forth.<sup>14</sup>

One important limitation of the subgroup review is that any subgroups that were completely excluded from existing randomized studies cannot be a part of published subgroup analyses using randomized study data. Another limitation of the subgroup review is its dependence upon what has been

---

<sup>12</sup>It is important to distinguish prognostic factors from correlates of the effect of a treatment. A patient's prognosis refers to his or her expected outcome in the absence of treatment; the issue of how much he or she might be helped by a particular treatment is a separate one. In some instances, patients with especially good prognoses might have the best chance of benefiting from a particular treatment. In others, those with poor prognoses might benefit most.

<sup>13</sup>A third approach would be to analyze a relevant data base or to use information from existing data base analyses. In chapter 4 of this report, treatment effects estimated in data base analyses are used for the purpose of filling gaps in randomized studies' coverage. This logically follows the assessment of the internal validity of data base analyses, presented in chapter 3.

<sup>14</sup>The subgroup review was discussed by Hyman (1955) and Cochran (1965). It has been explicated as the "elaboration model" (see Lazarsfeld, Pasanella, and Rosenberg (1972) and Rosenberg (1968)). The term "specification" has been used when different subgroups exhibit different relationships between the independent and dependent variables (i.e., when the relationship between the treatment and the outcome is different for different subgroups).



analyzed and reported. Existing subgroup analyses included in published reports may not cover all patient characteristics known to be underrepresented in randomized studies. For example, if it is known that blacks are underrepresented in existing randomized studies, one would look for studies that “break out” black patients, separately presenting the treatment’s effect for them. But the needed “break-out” or subgroup analysis may not be available. The reviewer could conduct a secondary analysis of randomized study data, but this may only rarely prove feasible. When the needed subgroup was not “broken out” in published reports, a meta-analysis technique could provide an alternative solution: As described below, meta-analysis can provide information on whether overrepresentation or underrepresentation of specific patient groups actually produces a different treatment effect.

Using meta-analyses to assess impact of nonrepresentativeness. When subgroup data on treatment effects have not been broken out by, for example, race or age, the reviewer can turn to a meta-analysis technique: checking for patterns in reported treatment effects across randomized studies that have been categorized according to the percentage of the underrepresented group (for example, the percentage of black or elderly patients) in their patient pools (see Devine and Cook, 1983; Light and Pillemer, 1984; Cordray, 1990b). Of course, in order to rule out the possibility that variability among findings derives purely from sampling error, meta-analysts have devised heterogeneity tests (see Hedges and Olkin, 1985). After conducting heterogeneity tests, the reviewer can test for patterned differences in the observed treatment effect across primary studies. Recently, sophisticated approaches such as hierarchical multiple regression have been applied in meta-analyses designed to distinguish (1) what portion of cross-study differences in observed treatment effects is attributable to varied patient characteristics, and (2) what portion of those differences is attributable to other sources (see Lipsey, 1992.)<sup>15</sup>

As noted above, these methods are useful for patient groups that are underrepresented and overrepresented in various randomized studies, but cannot address the issue of whether treatment effects differ in patient groups that have been totally excluded from all existing randomized studies.

---

<sup>15</sup>One cautionary note: Mosteller (1990a, p. 229) points out that such analyses “are ordinarily exploratory data analyses, and usually cannot be solidly confirmed by the same data that suggest the hypotheses.” This suggests that the reviewer using this meta-analysis technique should distinguish between (1) previously known indicators for which a hypothesis has been stated before the analysis, and (2) newly identified indicators uncovered as part of the current analysis.

## Assessing Representativeness Using Levels of Outcomes

### Part 1: Comparing Patient Outcome Levels

A somewhat lesser known two-part approach to assessing representativeness involves comparing patient outcome levels (for example, comparing the percentage of patients who survived for 5 years) across studies.

Perhaps the most notable example of outcomes comparison is the Coronary Artery Surgery Study (widely known as CASS), which involved both a randomized study and a follow-up of patients who declined to participate in the randomized study (see K. Davis, 1988). Patient outcomes in these two groups were compared in order to demonstrate the generalizability of the randomized trial. Survival curves for randomized and nonrandomized patients receiving each treatment were “virtually superimposable.” In other words, a demonstration of “no difference” between outcome levels observed for patients in a randomized study and for other patients receiving the same treatments was used to support a claim of generalizability.

Conversely, differences in outcome levels have been used to suggest a likely lack of generalizability. Kramer and Shapiro (1984) point to a study of portacaval-shunt surgery (Garceau, Donaldson, and O'Hara, 1964) in which survival of both the experimental and the control groups proved to be considerably higher than survival for other eligible patients who did not participate in the study. They argue that while this randomized study showed a zero-level treatment effect, its results may not be generalizable to the higher risk patients who did not participate.

Data demonstrating large differences in levels of outcomes have been obtained in a number of ways. For example, a follow-up of a Swedish primary prevention study used death records to compare outcomes of study participants and nonparticipants (Wilhelmsen et al., 1976). The nonparticipants consisted of those who had been asked to participate but had refused or not appeared at the study center. The comparison showed that during the follow-up period, mortality was three times higher among the nonparticipants than the participants.

Outcomes comparisons have been advocated by Remington (1989, p. I-67) as follows: “[A]n important guide to the representativeness of patients participating in [a] trial is the overall mortality rates in the group . . . .”

Certainly, many of the data bases used to compare baseline characteristics could also be used to compare patient outcome levels. Differences in outcomes for randomized-study participants and nonparticipants are



clearly interpretable when the magnitude of such differences dwarfs the magnitude of reported differences attributed to alternative treatments. The logic behind these outcome comparisons also seems clear when there is no difference between patient outcomes in the randomized study and the outcomes of other patients who received the same treatments. But a moderate difference in outcomes requires the analyst to account for potentially confounding factors, such as the influence of slightly different treatments (or levels of care) or the influence of treatment assignment bias on the outcomes of data base patients.<sup>16</sup>

Table 2.1 is presented as a preliminary guide to interpreting patterns of outcomes across studies and treatment groups. That is, table 2.1 distinguishes patterns that signal nonrepresentativeness from other patterns resulting from other sources.

For example, the center cell of table 2.1 shows one instance in which it appears that the patients in a randomized study are not representative. As indicated in the column and row headings for the center cell, this would occur if (1) an average outcome, such as survival, calculated for those patients in a randomized study who received a new treatment was lower than the corresponding average outcome calculated for data base patients receiving that same treatment, and (2) average survival for those patients in the randomized study who received usual care was also lower than for data base patients who received usual care. A similar situation applies for the upper left cell of table 2.1. Here, the participants in the randomized study have better survival than data base patients do. The pattern of better survival holds both when (1) the comparison is made using just patients who received the new treatment, and (2) it is made using only patients receiving usual care.<sup>17</sup>

<sup>16</sup>Of course, the problem does not exist if the new treatment has not yet been introduced into community medical practice. In such an instance, outcomes for control groups receiving usual care in trials may be compared to outcomes for all patients in the data base. Sometimes this situation can be simulated by focusing the comparison on data base patients in geographic areas where the new treatment has not yet been introduced.

<sup>17</sup>The patterns described in table 2.1 are limited to mean (average) outcome levels. But when dealing with a continuous outcome variable (e.g., years of survival), the standard deviation of outcomes for patients in a randomized study can be compared to the standard deviation for other patients, perhaps using a ratio. Logically, the standard deviation for an outcome variable would be high if the randomized study included both patients with extremely good prognoses and those with extremely poor prognoses. Conversely, if the standard deviation for patient outcomes in a randomized study is substantially lower than that observed for patients in a medical practice data base, patients in the randomized study may be a relatively narrow subset of the kinds of patients seen in medical practice. The meaning of the pattern is clearest when differences between the randomized study and the patient population dwarf differences between treatment groups in a single study.

**Table 2.1: Logic of Patient Outcome Comparisons: Patterns Signaling Nongeneralizability of Randomized Studies**

Outcomes for patients receiving usual care	Outcomes for patients receiving new treatment		
	$\bar{X}_t > \hat{\mu}_t$	$\bar{X}_t < \hat{\mu}_t$	$\bar{X}_t = \hat{\mu}_t$
$\bar{X}_{uc} > \hat{\mu}_{uc}$	<b>Patients in randomized study may not be representative</b>	Comparison bias likely in data base analysis	Combination of problems
$\bar{X}_{uc} < \hat{\mu}_{uc}$	Comparison bias likely in data base analysis	<b>Patients in randomized study may not be representative</b>	Combination of problems
$\bar{X}_{uc} = \hat{\mu}_{uc}$	Treatment implementation may differ: data base versus randomized study	Treatment implementation may differ: data base versus randomized study	Convergence

$\bar{X}_t$  = mean outcome for the treatment group of a randomized study

$\bar{X}_{uc}$  = mean outcome for the control group of a randomized study (in which patients were assigned to "usual care")

$\hat{\mu}_t$  = mean outcome for data base patients who received the new treatment

$\hat{\mu}_{uc}$  = mean outcome for data base patients who received usual care

Note: This table is a guide to distinguishing nonrepresentativeness of patients in randomized studies from two other problems: (1) comparison bias in data base analyses and (2) treatment implementation differences between randomized studies and data bases. The logic of the table assumes that any other potential problems were ruled out.

In other words, the key pattern is: Randomized study patients show a consistently higher—or a consistently lower—outcome level than do patients in the relevant population. This signals that patients in the randomized study may represent only the better—or only the worse—prognosis patients.

Of course, even if outcome levels indicate that patients with better prognoses (or those with worse prognoses) participated in a randomized study, the question remains as to whether the treatment effect differs across patients according to prognosis.

## Part 2: Linking Outcome Levels to Effect Sizes

To investigate whether the treatment effect differs across patients according to their prognoses, the reviewer can examine patient outcome levels (for example, survival rates) across various randomized studies, checking for linkage to effect sizes. Patients in some randomized studies may exhibit average survival levels above those for other randomized



studies. This pattern may occur both within the control group receiving usual care and within the new treatment group. Similarly, patients in other randomized studies may exhibit below-average survival levels. If such differences are found and if observed treatment effects vary beyond a level attributable to chance, questions such as the following can be asked:

- Are the highest treatment effects observed in the randomized studies that have captured patients with the best prognoses (for example, the longest surviving patients), while the lowest effects (perhaps zero-level or even negative) are observed in randomized studies conducted on patients with such poor prognoses that they may have been beyond help?
- Or did the reverse occur: Do the randomized studies whose patients have uniformly excellent outcomes (and hence "little room for improvement") report little benefit from the treatment?

Devine and Cook (1983) conducted a somewhat similar type of analysis as part of a meta-analysis of studies of psychoeducational interventions and length of postsurgical hospital stay. Here, the desired outcome was a shorter hospital stay. Devine and Cook found that over all studies, the intervention reduced hospital stays. But in those situations where patients already had a short average hospital stay (as was the case for the most recent studies), the intervention did not further reduce it.

## Summary of Task 1: Steps in Assessing the Generalizability of Existing Randomized Studies

The variety of approaches to assessing generalizability suggests that while each has some weakness, if all were applied to the same randomized study or set of randomized studies, the combined results should indicate either: (1) the nature and level of representativeness or nonrepresentativeness of patients participating in randomized studies, or (2) uncertainty about patient representativeness—and hence about generalizability of results—because of an inability to fully apply the foregoing assessments.

In other words, the foregoing assessments should allow the cross design investigator to complete the first task: assessing the generalizability of each randomized study's results. The specific steps that the investigator should follow are shown in table 2.2.

**Chapter 2**  
**Methods for Assessing Randomized Studies'**  
**Generalizability to the Patient Population**  
**(Task 1)**

**Table 2.2: Assessing Generalizability of Randomized Study Results to the Patient Population: Four Steps**

Target of assessment	Steps	Assessment methods for conducting each step
Methods of patient selection	<p><b>Step 1.</b> Assess every phase of the patient selection process for each randomized study.</p> <p><b>Step 2.</b> Assess likely representativeness of patients in each randomized study, using information from step 1 (all phases).</p>	<p>Assess each of the following for its likely impact on patient representativeness:</p> <ul style="list-style-type: none"> <li>—Patient selection criteria established: phase 1 (see T.C. Chalmers et al., 1981; McPeck, 1987).</li> <li>—Patient recruitment: phase 2 (see Pocock, 1983).</li> <li>—Exclusions, rejections of individual patients who met initial criteria: phase 3 (see T.C. Chalmers, et al., 1981; Remington, 1989).</li> <li>—Willingness of patients and physicians to participate: phase 4 (see Edlund, et al., 1985; Schooler, 1980).</li> </ul> <p>Judge likely representativeness given patient selection process: a serious problem in even one phase or small problems in multiple phases can threaten representativeness.</p>
Achieved representation	<p><b>Step 3.</b> Assess empirical non-representativeness on patient baseline characteristics and. . .</p> <p>. . .linkage to differential treatment effects.</p> <p><b>Step 4.</b> Assess empirical non-representativeness using patient outcome levels and. . .</p> <p>. . .linkage to differential effects.</p>	<p>Compare baseline characteristics of patients in the randomized study to those of patients in data base that approximates the full range of patients. Include characteristics related to treatment effectiveness (if any are known in advance), prognostic factors, and demographic variables (see Wilhelmsen et al., 1976; Steinhorn, et al., 1983).</p> <p>Conduct one or both of the following:</p> <ul style="list-style-type: none"> <li>—a subgroup review of randomized studies (see Cochran, 1965; Hyman, 1955; Lazarsfeld et al., 1972; Rosenberg, 1968).</li> <li>—a meta-analytic comparison of results across randomized studies that differentially represent patient groups (see Devine and Cook, 1983; Light and Pillemer, 1984; Cordray, 1990b; Lipsey, 1992).</li> </ul> <p>Compare outcome levels of patients in randomized studies to those of patients in a data base that approximates the full range of patients (K. Davis, 1988; Wilhelmsen et al., 1976; Remington, 1989; S. Davis et al., 1985; Kramer and Shapiro, 1984).</p> <p>Conduct one or both of the following:</p> <ul style="list-style-type: none"> <li>—an analytic subgroup review, checking whether treatment effects differ for subgroups with better (or worse) overall outcomes.</li> <li>—a meta-analytic comparison of results across randomized studies that vary in terms of outcome levels (Devine and Cook, 1983).</li> </ul>



Once steps 1 through 4 have been completed for each randomized study, the cross design investigator combines this information to reach a conclusion about the nature and extent of any generalizability problems associated with each randomized study. Alternatively, the investigator may determine that there is uncertainty concerning the generalizability of a randomized study's results because of, for example, a lack of information needed to complete the assessments.

To our knowledge, these assessment steps have not been presented previously as a set. Thus, a number of potential technical issues might be raised about their use in combination; for example, whether significance tests conducted in one step would affect the form of the tests to be conducted in another step or whether an investigator should complete each step without reference to the results of other steps. These, and doubtless other technical issues, need to be explored by future analysts.

To complete assessment of generalizability of existing randomized studies, the investigator considers all randomized studies in the existing set with respect to their combined coverage of the relevant patient population. This assessment of all randomized studies taken together guides the investigator's decision about the need for a cross design synthesis, as opposed, for example, to a conventional meta-analysis of randomized studies' results. That is, the investigator decides whether or not there is a need to combine results of randomized studies with information from analyses of observational data bases that more fully represent the relevant patient population.

If the investigator does decide upon a cross design synthesis, then the results of the assessments conducted separately for each randomized study will provide the information needed to:

- adjust individual randomized study results, standardizing them to better reflect the composition of the relevant patient population;
- develop an appropriate framework for the synthesis; and
- combine the results of each randomized study with the results of other randomized studies and with other kinds of studies in a way that is appropriate, given the certainty associated with each randomized study's results.

These tasks are discussed in detail in chapter 4.

# Methods for Assessing Data Base Analyses for Comparison Bias (Task 2)

Today, patient records are routinely stored in computer data bases maintained by hospitals and insurance companies. In addition, certain “specialty” data bases have been compiled to develop clinical information on a particular disease (for example, cancer). Many administrative and clinical data bases come close to encompassing the full range of patients seen in medical practice.<sup>1</sup> Several of these “medical practice” data bases have been stripped of patient identifiers and made available for use by researchers.

Data base analyses have now been conducted to estimate treatment effects. These estimates often provide much fuller coverage of the patient population than existing randomized studies do. When this is the case, data base analyses represent a potentially useful addition to the results of randomized studies. But despite greater coverage of the patient population, data base analyses may not be suitable for synthesis with randomized study results. As discussed in chapter 1, the problem is one of internal validity or “comparison bias.”

Briefly, in medical practice, patients and their physicians freely “assign,” or choose from, alternative treatments according to their preferences. Patients who choose one treatment may be very different from those who choose another. If the patients choosing the new treatment have better (or worse) prognoses than those receiving usual care, it follows that outcomes for the two patient groups would differ even without a difference in treatments. In this instance, comparing the outcomes of patients receiving the two treatments would yield an invalid, biased treatment effect. This problem of “comparison bias” can mar the results of any data base analysis.<sup>2</sup>

Only randomization ensures unbiased treatment assignment. However, a data base analyst can define and adjust comparison groups to minimize

---

<sup>1</sup>Many data bases cover only a specific segment of the population; for example, a Medicare data base covers only elderly patients, but it is quite comprehensive for this group. Certain other data bases may appear to be comprehensive, but in fact fail to capture the full population. A case in point is the data base of AIDS cases maintained by the Centers for Disease Control. The sizable “undercount” of AIDS cases in that data base was estimated by GAO (1989b). Thus, it may be appropriate for an investigator to assess the generalizability of a data base, using methods such as those outlined in chapter 2 of this report.

<sup>2</sup>In this report, the primary term used to refer to this problem is “comparison bias.” Terms used to refer to specific aspects of the problem include “treatment assignment bias” and “imbalanced comparison groups.” Other terms used in the literature (but not in this report) include “selection bias” (Byar, 1980) and “nonignorable treatment assignment” (Rubin, 1978). The term “selection bias” is potentially confusing because it has been used to refer to both the problem discussed here and lack of generalizability. “Nonignorable treatment assignment” is more specific, but may be difficult for nontechnical readers.



imbalance. A thorough assessment of the data base analyst's methods of comparison and adjustment will help suggest the likely dimensions of comparison bias in a particular analysis. In addition, empirical assessment of the degree of balance actually achieved in that analysis can provide further information on the nature and degree of comparison bias.

Multiple assessments of comparison bias can, when taken together, indicate and enhance the usefulness of data base results. But the challenges involved should not be understated: Even when no imbalance in the comparison groups is found, hidden imbalances may exist. And where several imbalances are noted and taken into account, other imbalances may go undetected. To meet this challenge, we emphasize that multiple assessment approaches should be used in combination. These include, among others, sensitivity analyses that address the question of whether undetected imbalances represent a threat of potentially serious proportions.

For the cross design investigator, the specific purposes of assessing comparison bias are to:

- determine whether treatment effects estimated in existing data base analyses are sufficiently free of comparison bias to be suitable for combining with randomized study results;
- provide, when appropriate, a basis for the secondary adjustment of data base results to further minimize comparison bias; and
- help estimate the level of uncertainty regarding hidden comparison bias in the adjusted data base results.

Previously reported assessments of comparison bias fall into two major categories: (1) judgmental assessments of the methods of comparison and adjustment used by the data base analyst; and (2) empirical assessments of the achieved balance of the comparison groups following adjustment by the data base analyst.

---

Assessments of the methods a data base analyst used to minimize comparison bias must be based largely on the judgment of the reviewer. Three examples of judgmental assessments are provided by:

- Wortman and Bryant's (1985, p. 292) review of school desegregation and academic achievement studies, which included a qualitative a priori judgment about whether or not each study had "selection problems."

---

## Judgmental Assessments of Methods Used by Data Base Analysts

- Longnecker et al.'s (1988, p. 653) meta-analysis of nonrandomized studies on alcohol consumption and breast cancer risk, which included a rating-score instrument with the following item: "In the analyses, did the authors control for potential confounding by classic breast cancer risk factors in addition to age?"<sup>3</sup>
- Steinberg et al.'s (1991, p. 1986) meta-analysis of case-control studies on estrogen replacement and breast cancer risk, in which three epidemiologists judgmentally assigned scores, using the following criterion, among others: "appropriate potential confounding factors were ascertained...and...the analyses were adjusted for these confounders."

Assessments of the methods that the data base analyst used to minimize comparison bias can proceed in two phases:

- In phase 1, the reviewer assesses the data base analyst's choice of comparison groups.
- In phase 2, the reviewer assesses the data base analyst's adjustments of those groups.

---

## Phase 1: Assessing the Analyst's Choice of Comparison Groups

Most data base analysts have chosen to use treatment groups as comparison groups. That is, they have compared the outcomes of patients who received a new treatment to the outcomes of those who received usual care. Alternatively, however, the analyst may have chosen "natural cohorts" as comparison groups.<sup>4</sup> The assessment question is: Which set of comparison groups is the better choice for minimizing comparison bias—treatment groups or natural cohorts? The answer depends upon the particular situation.

### Background: Treatment Group Comparison

Usually, data base analysts have estimated medical treatment effects by comparing patient outcomes across treatment groups. A variety of statistics may be used, such as the difference between each group's average outcome or the relative risk of a negative outcome following the

---

<sup>3</sup>The risk factors adjusted for in these breast cancer studies include the patient's age at first childbirth and her body mass, among others (see, e.g., Schatzkin et al., 1987).

<sup>4</sup>This lesser known approach is described below, following background on the more common treatment groups comparison.



new treatment as opposed to usual care or the correlation between treatment and outcome.<sup>5</sup>

A well-known example of the treatment group approach is the comparison of mortality following open prostatectomy to mortality following a newer approach—transurethral resection of the prostate. Two such studies (one based on a Canadian data set, the other based on a Danish data set) found higher mortality following the newer treatment (see Roos et al., 1989, and Andersen et al., 1990). In both cases, relative risk was used to estimate the treatment effect.<sup>6</sup>

Another version of treatment group comparison involves not just two treatment groups but a treatment variable that takes on many different values. A case in point consists of studies of the effect of the timing of breast cancer surgery in the premenopausal patient's menstrual cycle (that is, surgery conducted on day 1 through day 28-32). Two analyses—one using data from a previous Los Angeles community study, the other using data from Guy's Hospital in London—showed that outcomes such as disease recurrence and mortality varied depending on the day of the menstrual cycle when surgery was performed (Hrushesky et al., 1989; Badwe et al., 1991).<sup>7</sup>

#### Background: Natural Cohort Comparison

"Natural cohorts" are pre-existing patient groups (or naturally occurring patient groups) that differ in terms of the relative numbers of patients receiving the new treatment and usual care. The analyst compares net outcomes across these naturally occurring groups.<sup>8</sup>

For example, if patients diagnosed in each successive 1-year or 2-year interval were markedly more likely to receive a new treatment, these annual or biannual diagnostic groups would constitute natural cohorts. An analyst could compare outcomes for all patients diagnosed in a given year to outcomes for all patients diagnosed in each subsequent year, thereby

---

<sup>5</sup>A variant on treatment group comparison uses change scores as the outcome measure. One example is cognitive pretests and posttests in an Alzheimer's drug treatment trial (see T. Thompson et al., 1990). This approach is precluded in those medical studies where the outcome measures (such as disease recurrence or death) are "one time only" and occur after treatment.

<sup>6</sup>The relative risks were calculated using adjustments to balance the comparison groups. These adjustments are described in a subsequent section of this chapter.

<sup>7</sup>A third study (Senie et al., 1991) has also been reported. Findings from these small-sample studies do conflict, especially when cutting points are used to divide the month into two intervals (one "risky" and one "safe"). The riskiest time in all three studies appears to be around the 4th to 12th day from the beginning of the last menstrual cycle.

<sup>8</sup>Such analyses have sometimes been referred to as "natural experiments."

gaining an estimate of the overall impact that increased use of the new treatment has had on patient outcomes.

As delineated by Moffitt (1991), an analysis of this type may be advisable whenever it is possible to identify a variable that is both (1) related to the likelihood of receiving a new treatment, and (2) unrelated to patient prognoses. The importance of the second requirement is evident when one considers that comparison bias can occur in natural cohorts. For example, if patients experienced more severe cases of a disease in each successive year, then net comparisons of outcomes across annual cohorts would not isolate the impact of changes in treatment over the years. Indeed, one would not know whether differences in outcomes resulted from increasing use of a new treatment or from changes in the severity of the disease.

An example of a natural cohort comparison is the analysis of chemotherapy and breast cancer survival that GAO conducted, using the SEER data base (GAO, 1989a). This analysis focused solely on premenopausal, node-positive breast cancer patients and compared survival across natural cohorts defined by year of diagnosis. That is, survival for patients diagnosed in 1975 was compared to survival for patients diagnosed in 1976, and so on through the early 1980s. These annual cohorts of breast cancer patients apparently met both criteria discussed above:

- First, differing proportions of the node-positive patients in each cohort received chemotherapy; that is, only 23 percent of the 1975 cohort received chemotherapy, but 45 percent of the 1976 cohort received it, as did ever larger proportions of subsequent cohorts through 1981 (when 66 percent of patients received it).
- Second, as measured by such prognostic factors as size of tumor, the prognoses of the node-positive patients—their expected outcomes absent new treatment—remained steady across the cohorts.

The results of this breast cancer study showed that, despite more widespread use of chemotherapy over the years, there was no detectable improvement in the overall survival of premenopausal node-positive breast cancer patients: 72 percent of the entire 1975 cohort survived 5 years, as did 71 percent of the entire 1976 cohort and 72 percent of the 1981 cohort. A difference in outcomes across the natural cohorts would have indicated a positive (or negative) effect of the new treatment.



Another study (Wennberg et al., 1989) provides an example of geographic locations as natural cohorts. Boston and New Haven differ in terms of overall hospital usage: Boston has both greater numbers of patients admitted to hospitals and longer lengths of hospital stay. This difference in treatment choices (that is, hospital use) appeared to stem from differences in medical opinions on whether certain diseases actually require hospitalization, rather than from differences in the incidence or severity of disease (which appeared to be similar in the two cities).<sup>9</sup>

Using adjustments to equate sex, race, and age in the two cities, Wennberg et al. compared hospital use rates and mortality rates. The results showed that the adjusted mortality rates were the same in both cities: Lower hospital usage in New Haven apparently did not have a negative effect on patient survival. A difference in adjusted mortality rates would have indicated the presence of an effect.

Where there are two natural cohorts, the size of the treatment effect can be estimated in two steps. For example:

- Step 1: The difference in outcomes observed for the two natural cohorts is calculated by subtraction.
- Step 2: The difference in outcomes is divided by the difference between the proportion of patients receiving the new treatment in each natural cohort.<sup>10</sup>

Step 1 (subtraction) is the same procedure that is often conducted for a treatment group comparison. Step 2 is an additional procedure that is necessary for a natural cohort comparison. Step 2 is not needed in a treatment group comparison because 100 percent of one group received the new treatment and zero percent of the other group received it.

---

<sup>9</sup>Specifically, analysis of the Medicare data base showed that the hospital use rates in Boston were higher for those "medical conditions, such as pneumonia, gastroenteritis, and chronic obstructive lung disease, for which there is little consensus about the need for hospitalization .... By contrast, discharge rates involving myocardial infarction, stroke, and gastrointestinal hemorrhage ... were virtually the same in the two communities. For these ... conditions, which are characterized by professional consensus on the need for hospitalization, the hospitalization rates are more closely related to the incidence rates of the disease." (Wennberg et al., 1989, p. 1168.)

<sup>10</sup>That is: Estimated effect =  $(\text{MEAN}(Y_1) - \text{MEAN}(Y_2)) / (p_1 - p_2)$

Where  $Y_1$  refers to outcomes in natural cohort 1,  $Y_2$  refers to outcomes in natural cohort 2,  $p_1$  refers to the proportion of cohort 1 that received the new treatment, and  $p_2$  refers to the proportion of cohort 2 that received the new treatment.

In situations where  $p_1 = 1$  and  $p_2 = 0$ , this formula reduces to the simple difference in means formula, which is appropriate for a treatment group comparison.

Step 2 is required in a natural cohort comparison in order to correct for the dilution of the observed difference. The dilution occurs because in the cohort where the new treatment predominates, the percentage receiving it is less than 100 percent; in the other cohort, the lower percentage receiving the new treatment is greater than zero. For example, if the proportions of each cohort receiving the new treatment are 0.90 and 0.20, then the denominator in step 2 would be: 0.9 minus 0.2, or 0.7. Dividing the observed difference by 0.7 would inflate it to the appropriate level. Thus, an observed difference of 10 percent would be inflated to 14 percent (since 0.1 divided by 0.7 is 0.14). Angrist (1990) cites Wald (1940) as the source of such estimates, referring to them as “Wald estimates.”<sup>11</sup>

In econometrics, the approach we term “natural cohort comparison” is referred to as “instrumental variables estimation” or as the “identifying variables solution.” If multiple variables define natural cohorts (groups more and less likely to receive the treatment), econometricians carry out a two-stage least-squares regression. In the first stage, the identifying variables are used to predict treatment assignment. In the second stage, the predicted treatment assignment is used to predict outcome.<sup>12</sup> The key, however, is to find at least one “legitimate” identifying variable that meets both requirements discussed earlier (that is, one set of natural cohorts).

### Assessing Choice of Treatment Group Versus Natural Cohort Comparison

The task of the reviewer assessing comparison bias is to determine whether, in a given data base analysis, it is more appropriate to use a treatment group comparison or a natural cohort comparison. The answer depends primarily upon which approach provides the more balanced comparison:

- A comparison of outcomes across treatment groups (such as the prostatectomy analyses discussed on page 55) is unbiased provided that the different treatment groups are balanced (that is, provided that patients in the different groups have comparable prognoses or the same expected

<sup>11</sup>Where covariances are used to define an effect for natural cohorts, there are two analogous steps. In the first step, the analyst calculates the covariance of (1) the variable defining the cohorts and (2) the outcome variable. In the second step, the analyst divides the covariance calculated in first step by the covariance of (1) the variable defining the cohorts and (2) the treatment variable (see a recent econometrics text, such as Wallace and Silver, 1988). In the breast cancer example, step 1 would consist of calculating the covariance of the diagnostic year and patient survival. The denominator, calculated in step 2, would be the covariance of the diagnostic year and the receipt of chemotherapy.

<sup>12</sup>In statistical terms, “first regress...X on Z, then regress...Y on predicted X” (Wallace and Silver, p. 255). Here, X refers to the treatment variable; Z refers to the variable defining the “natural cohorts” or to a set of such variables; and Y refers to the outcome variable. Thus, in the chemotherapy/breast cancer example discussed above, the analyst would predict receipt of chemotherapy from annual diagnostic cohort and perhaps also geographic location; the analyst would then predict survival from predicted receipt of chemotherapy.



outcomes following usual care). Said another way, a treatment group comparison is valid to the extent that an individual patient's assignment to a particular treatment alternative is not linked to the patient's prognosis.<sup>13</sup>

- A comparison of outcomes across “natural cohorts” that differ in terms of the prevalence of a new treatment (such as in the analysis of chemotherapy’s effect for breast cancer patients) represents a valid estimate of the treatment’s effect provided that the cohorts are balanced in terms of patient prognoses. A natural cohort comparison is valid to the extent that differences in the prevalence of a particular treatment across cohorts are not linked to differing patient prognoses in those cohorts.

Table 3.1 provides a guide to assessing an analyst’s choice of treatment group comparison versus natural cohort comparison.

Table 3.1: Assessing the Analyst’s Choice of Comparison Groups: Treatment Groups Versus Natural Cohorts

Prevalence of a treatment across different cohorts is linked to patient prognoses in those cohorts	Individual treatment decisions are based on or linked to patient prognosis	
	Yes	No
Yes	Situation 1: Neither is preferable; both may be invalid	Situation 2: Treatment groups are preferable
No	Situation 3: Natural cohorts are preferable	Situation 4: Either is valid; results should converge

Depending upon patterns in the underlying data, one or the other type of comparison may be more appropriate. If treatment groups are indeed comparable (situations 2 and 4), there is no need for the analyst to search for “natural cohorts.” If both types of comparisons are valid (situation 4) and data on natural cohorts are available, a natural cohort comparison may be used to confirm a treatment group comparison.

However, physicians often choose between alternative treatments based on the patient’s prognosis. For example, whether cancer patients are advised to undergo more or less aggressive therapies often depends upon the stage of their disease. This suggests that in certain instances treatment groups may be a poor choice for outcome comparisons.

The literature on “small-area variations” in medical practice (see Wennberg et al., 1988; 1989) shows that physician preferences can vary

<sup>13</sup>Or, in the instance of a change-score analysis, it is the expected change scores following usual care that must be balanced across treatment groups.

independently of patient prognoses. Given two alternative treatments for prostate disease, for example, one may be much more popular among physicians in certain geographic areas, while the other may be favored by other physicians in different areas. Physicians' preferences for particular treatments may be linked to differences in their training, philosophies (such as whether survival is more important than quality of life), or beliefs about the relative effectiveness of different treatments (see Wennberg et al., 1988).<sup>14</sup> Preferences may also be linked to the availability of technology and resources. Thus, a key question becomes: Do physician preferences for certain treatments differ geographically (for example, across urban and rural areas)—or across time?

The reviewer must also remain alert to the possibility that patients sometimes do differ in severity of disease across geographic areas or across diagnostic years. Diagnostic improvements in large cities could, for example, result in the detection of greater numbers of less severe cases in cities than in rural areas. A similar situation could occur over time.

Ideally, data on these factors—as they impact a specific treatment—would be brought to bear by reviewers assessing an analysis of that treatment's effect. Alternatively, the reviewer should draw upon a broad knowledge of patterns of medical practice regarding the disease in question.

---

## Phase 2: Assessing Adjustments the Analyst Made to Balance Comparison Groups

An assessment of methods used by the data base analyst should include a review of the adjustment procedures used to eliminate imbalances in the comparison groups. Because treatment groups are rarely equivalent, data base analysts have often made adjustments to increase the comparability or balance of the groups. For example, in comparing mortality following open prostatectomy to mortality following transurethral resection, Andersen et al. (1990) introduced adjustments for age and for comorbidities, which followed the model used by Roos et al. (1989). Essentially, these adjustments equated the treatment groups on measured prognostic factors (that is, covariates of the outcome absent treatment).

Similarly, adjustments may be needed to balance natural cohorts. We know of only one study where the analyst could use cohorts formed by

---

<sup>14</sup>A study of Maine urologists showed that some advocated prostatectomies for patients with prostate problems that had not resulted in chronic obstruction—regardless of symptom relief. Such physicians believed “that life expectancy is improved by avoiding the need for operation at a later date” (Wennberg et al., 1988, p. 3028). Other physicians disagreed; they believed that prostatectomy would be justified for such patients only if symptoms would be reduced and quality of life improved.



## Background: Overview of Adjustment Techniques

random assignment.<sup>15</sup> In all other instances, imbalance should be assessed and adjusted for.

In medical studies, the statistical techniques that have been used to equate treatment groups on prognostic factors (and could also be used to equate natural cohorts) include:

- matching (Cochran 1965; Cochran and Rubin, 1973);
- subclassification or “poststratification” (for example, separate examination of treatment effects for the healthiest stratum of patients) and adjustments based on, for example, weighted averages taken across strata (Cochran, 1965; Cochran and Rubin, 1973); and
- analysis of covariance (Cochran and G. Cox, 1957); partial correlation (DuBois, 1957), regression (Cochran and Rubin, 1973), the Cox proportional hazards model (D.R. Cox, 1972, 1975), and “propensity scores” (Rosenbaum and Rubin, 1983a).<sup>16</sup>

For reviews of the above techniques, see Rubin (1984) and Kish (1987).<sup>17</sup> Other techniques for obtaining unbiased comparisons include:

- selection modeling (Cain, 1975; Heckman and Hotz, 1989a, 1989b; Rindskopf, 1986); and
- structural equation models (see Rindskopf, 1981; Bentler, 1990; Joreskog, 1977; Goldberger and Duncan, 1973).<sup>18</sup>

Potential problems and complexities in the use of the various adjustment methods are discussed by Reichardt (1979). Adjustments can be risky when large imbalances on the measured covariates distinguish the groups or cohorts being compared. Cochran has advised analysts planning

<sup>15</sup>During the Vietnam war, random assignment of birthdays for the military draft lottery created natural cohorts; that is, groups of young men that differed in terms of the likelihood of military service. The incomes of men in these randomly formed cohorts were compared for the years following the Vietnam war (Angrist, 1990). It is interesting to note that an analysis based on randomly formed natural cohorts is comparable to a randomized study in which (1) there are numerous “crossovers” who obtain a treatment other than the one assigned to them, and (2) the comparison of patient outcomes is made on the basis of the treatment that was assigned (not the one eventually obtained).

<sup>16</sup>These statistical techniques may be more effective when used in combination; for example, subclassification might best be combined with other, multivariate adjustments (Lavori et al., 1983, pp. 1292, 1294). Similarly, a combination of matching and adjustments based on multiple regression has been demonstrated to be more effective than multiple regression alone (Cochran and Rubin, 1973).

<sup>17</sup>Adjusting outcomes using nonlinear models for binary outcomes is more complex (see Chang, Gelman, and Pagano, 1982).

<sup>18</sup>Structural equation models are not unrelated to selection models. Structural equation models can include prediction of treatment assignment as well as prediction of the effects of treatment received.

observational studies to “avoid treatment and control groups with large initial differences on confounding variables” (Rubin, 1984, p. 41, citing Cochran, 1965). Adjustments are also risky when there is measurement error in the covariates used to make adjustments.<sup>19</sup> Because of the complexities sometimes involved in adjustments, a methodological assessment of the techniques used should include a statistical expert’s opinion on whether various pitfalls were ruled out in the original data base analysis (see Kaplan and Berry, 1990, p. 113).

Selection modeling, which includes a variety of approaches, is of special interest here because it has been applied in deriving instrumental variable estimates (or “natural cohort comparisons,” as we have termed them in this report; see Moffitt, 1991). Specifically, in the second stage of the two-stage least-squares analysis, correlates of the outcome variable are entered as controls (see Rindskopf, 1986).

Controversy has surrounded selection modeling because of apparent claims by some advocates of this approach that unmeasured differences between comparison groups can be eliminated.<sup>20</sup> We believe that the requirement for balance is never eliminated, but merely shifts to the particular groups that are being compared. For example, with respect to the natural cohorts approach described earlier, unmeasured differences across the treatment groups become irrelevant; but unmeasured differences across the “natural cohorts” are potentially of concern.

### Assessing Whether Adjustments Covered a Sufficient Set of Prognostic Factors

If the set of prognostic factors that the data base analyst adjusted for is insufficient, then there will be remaining imbalances in the comparison groups and these will bias the results. The problem is essentially the same for adjustments to balance treatment groups and adjustments to balance natural cohorts. It exists regardless of the specific control or adjustment technique used. The major assessment task of the reviewer is thus to determine whether the data base analyst adjusted for a sufficient or complete set of prognostic factors.

<sup>19</sup>The controversial first evaluation of Project Head Start (the Ohio State-Westinghouse evaluation) is an instance where measurement error in a covariate may have combined with the imbalance in the comparison groups, thus threatening the validity of the adjustments on that covariate. For two very different views, see Barnow, Cain, and Goldberger (1980), and Kaplan and Berry (1990, citing Campbell and Erlebacher, 1970).

<sup>20</sup>For a brief discussion of this controversy, see Coyle, Boruch and Turner, 1991, which also notes that in spring 1991, the National Research Council of the National Academy of Sciences recommended that the National Science Foundation sponsor research into the empirical accuracy of estimates of program effects derived from selection model methods. See also, Holland, 1989; Moffitt, 1989; Heckman and Hotz, 1989c.



An example of the difference that hidden prognostic factors can make in balancing comparison groups is provided by two successive analyses conducted at the Health Care Financing Administration (Roper et al., 1988). These analyses compared survival following coronary artery bypass surgery with survival following balloon angioplasty. The initial analysis, which adjusted for covariates readily available in the data base, showed an apparent benefit for angioplasty:

"The ratio of the probability of dying over a period of up to two years after angioplasty to the probability of dying under similar conditions after bypass surgery is 0.7 according to the Cox proportional-hazards model ( $P=0.002$ ), after adjustment for age, sex, race, the incidence of selected high-risk comorbid conditions, and the use of either procedure as an emergency treatment of acute myocardial infarction." (pp. 1199-1200; emphasis added)

However, a further, heroic effort was made to supplement the HCFA data base with additional patient records that would allow further adjustments for treatment assignment bias. Specifically, the HCFA data base was supplemented by:

"a large number of preoperative clinical findings, obtainable from medical records through the PRO system...includ[ing] historical data, observations from physical examinations, and the results of laboratory and diagnostic tests." (p. 1199)

This second analysis showed that:

"If more clinical characteristics are taken into account, such as a previous bypass, the presence before an operation of atrial fibrillation and a depressed left ventricular ejection fraction—a total of 37 findings predictive of death after one or both procedures—the risks of death are virtually indistinguishable. Thus ... the probability of dying after the two procedures [is] nearly equal .... The patients at higher risk appear to be undergoing the more complex procedure—bypass." (p. 1200, emphasis added)<sup>21</sup>

How does a reviewer address the question of what might constitute a "sufficient" set of covariates in a particular adjustment? The Longnecker et al. (1988) rating-scale item cited earlier assumes a priori knowledge about classic breast cancer risk factors. In many instances, it would be reasonable to suppose that the investigator would be knowledgeable about important prognostic factors; however, in the absence of such knowledge, the investigator can refer to:

<sup>21</sup>An adjustment that reduces a treatment effect to zero is likely to be a successful one, but this is not necessarily the case. For example, the adjustment model may have increased imbalance on a hidden variable. Tests that help analysts assess the appropriateness of adjustments are discussed later in this chapter.

- Existing models of specific diseases, which have been published in the literature and which identify the importance of various prognostic factors and their interrelationships. One example in the area of heart disease is a model developed at Duke (Pryor et al., 1983; see also Hlatky et al., 1988), which specifies the interrelationships of demographic factors (such as patient sex, age), behavior (such as smoking), disease history (such as previous heart attacks), and various clinical and laboratory findings.
- Generic scales that address “case mix” issues relating to patient outcomes or that rate coexisting illnesses.<sup>22</sup> Examples would be (1) the APACHE classification system,<sup>23</sup> which has been shown to predict patient outcomes (Wagner, Knaus, and Draper, 1983), and (2) the Charlson comorbidities index (Charlson et al., 1987), which has been used to control for treatment assignment bias (see, for example, Roos et al., 1989).

Using the information from such models and scales as a standard, the reviewer can make a judgmental assessment of the completeness of the set of patient characteristics (covariates of outcome) used by the original data base analyst to detect imbalances in the comparison groups and adjust for them.

Of course, an omitted prognostic factor will bias results only if the comparison groups were, in fact, imbalanced following the adjustments that were made. This highlights the fact that, in addition to judgmental assessment of methods used by the data base analyst, empirical assessments of the results of those methods should be conducted. Empirical assessments are necessary to determine the achieved balance of the comparison groups.

## Empirical Assessments of Achieved Balance in Adjusted Comparison Groups

Despite their efforts to achieve balanced comparison groups, data base analysts are rarely 100-percent successful. As Mosteller et al. (1985, p. 109) have pointed out:

“Because data bases ordinarily contain information from patients whose treatment was chosen in an uncontrolled manner and delivered in an uncontrolled and poorly monitored fashion, groups receiving different treatments cannot be expected to be similar in prognosis. Attempts to use data to compare the effects of different treatments must

<sup>22</sup>The term “case mix” (Greenfield, 1989, p. 1143) has been used to summarize the combination of patient characteristics—including the severity of the illness, coexisting conditions, and other factors (such as age and functional status) that might affect a patient’s outcome regardless of treatment received. Hornbrook (1985) provides a review of six approaches to measuring case mix. For discussions of primary care measures of case mix, see White, 1991, and Weiner, 1991.

<sup>23</sup>Acute Physiology and Chronic Health Evaluation.



therefore use analytic devices to attempt to remove the effects of biases. Such devices are not entirely satisfactory....”

A similar argument could be made about balancing natural cohorts. The analyst’s degree of success in balancing comparison groups (whether these are treatment groups or natural cohorts) varies from study to study.

To assess the success of the comparison and adjustment methods that were used, data base analysts themselves—and others performing secondary analyses—have used numerous empirical tests involving data manipulation. These include no-difference tests, sensitivity analyses, and goodness-of-fit tests. Other assessments that compare results across studies have been developed primarily by meta-analysts and by medical researchers studying coronary artery disease.

Assessments of the data base analyst’s success in balancing comparison groups fall into three main categories:

- review of the empirical tests of achieved balance provided by the data base analyst;
- secondary analyses to test for achieved balance; and
- comparisons of the results of the data base analysis to other studies’ results.

## Reviewing Empirical Tests Provided by the Data Base Analyst

The literature shows that many data base analyses have included empirical tests of whether balanced comparison groups were actually achieved. These tests include: (1) tests of “no difference” in comparison groups, (2) sensitivity analyses, and (3) goodness-of-fit tests. Reported results of these tests can provide crucial input to a reviewer’s assessment of whether postadjustment imbalances in the comparison groups continue to bias the estimated treatment effect.

## Tests of “No Difference” in Comparison Groups

Demonstration of “no difference” in the baseline characteristics of comparison groups has been an important element in the evidence marshaled to support causal interpretations (see Yeaton, 1990). A review of studies published in the New England Journal of Medicine (Lavori et al., 1983) indicates that the majority of nonrandomized studies compared alternative treatment groups on several prognostic factors; each study

showed that its comparison groups were essentially the same on these variables.<sup>24</sup> Similarly, our breast cancer study (GAO, 1989a) reported that essentially no differences distinguished the annual cohorts for certain prognostic factors (such as size of tumor).

One way to expand data available for checking imbalances in the patient comparison groups is to contact the original data base analyst. Basic computer runs comparing baseline characteristics are likely to have been made.<sup>25</sup> If imbalances are found on patient characteristics other than prognostic factors (such as demographics), these differences cannot be assumed to bias results. The estimated treatment effect would be distorted only if the imbalances are linked to persistent differences in patient prognoses.<sup>26</sup>

An ingenious technique for no-difference testing (developed by Rubin, 1991) compares the adjusted groups on a patient outcome measure—one that is different than the one used in the main analysis. Indeed, unlike the outcome criterion of the main analysis, the outcome measure for this test should be unrelated to the effect of the treatment. The logic of this test is as follows: If the comparison groups are balanced (that is, have equivalent prognoses except for the effect of the treatment), then the two groups should be the same on the test outcome measure.

This no-difference test was used for an analysis of a 1981-85 Medicare data set, in which “the effects of switching from a name-brand to a generic drug” were examined (Rubin, 1991, p. 1213). The generic form of thioridazine became available in March 1983 (the intervention date). During the following months, patients had the option of switching from a name-brand version of this drug to the generic form. To analyze the effect

<sup>24</sup>Statisticians have pointed out that significance tests are not relevant when comparing baseline characteristics to assess imbalances that could lead to treatment assignment bias. Specifically, one should not “interpret ... a nonsignificant statistical difference as constituting sufficient evidence that the groups were not substantially different” (Lavori et al., 1983, p. 1291). Certainly, the importance of imbalances in groups receiving alternate treatments should not be determined on the basis of whether or not the imbalances occurred by chance alone (Altman and Doré, 1990). In randomized studies, chance differences between treatment groups are not considered irrelevant; hence, they should not be considered irrelevant in data base analyses.

<sup>25</sup>Ideally, the reviewer would seek to validate these results, perhaps by inquiring as to who performed the runs and checking whether they were reviewed by the principal investigator.

<sup>26</sup>For example, suppose that, on average, patients who received a new treatment were younger than those who received usual care and that an imbalance in the ages of the patients in the comparison groups remained after the analyst adjusted for other variables. Would there be a resultant bias in the observed treatment effect, calculated by comparing outcomes for these two groups? The answer is “yes” if the difference in patient ages for the comparison groups is related to a difference in the prognoses of these groups—even after adjustments for various prognostic factors.



of switching, a treatment group comparison was made. Specifically, matched pairs of nonswitchers and switchers were carefully developed. Each pair was characterized by a switch date. The main analysis focused on outcomes after the switch date for each pair. Patients who continued to use the name-brand version of the drug (nonswitchers) were compared to paired patients who switched to a generic substitute. The nonswitchers appeared to fare better than switchers.<sup>27</sup>

To check this result, several tests were performed including a no-difference test that involved patient outcomes during the interim between the intervention date and the switch date for each pair. If the pairing and the adjustments used in the main analysis had truly equated switchers and nonswitchers, then their outcomes during this interim period should be very similar.

## Sensitivity Analyses

Sensitivity analyses are essentially simulations of outcomes under alternative conditions. Such analyses can be used to:

“examine the sensitivity of estimates to assumptions about unobserved covariates .... If estimates are relatively insensitive to plausible variations in assumptions about unobserved covariates, then a causal interpretation is more defensible” (Rosenbaum, 1984, p. 42, citing Cornfield et al., 1959, Rubin, 1978, and Rosenbaum and Rubin, 1983b).

An example of a sensitivity analysis is presented by Rosenbaum and Rubin (1983b) in a nonrandomized study of coronary artery disease. Briefly, following adjustments, proportions of patients experiencing functional improvement were estimated for medical therapy (0.36) and for surgery (0.67). Rosenbaum and Rubin then examined the sensitivity of these estimates to a hypothetical unmeasured covariate related to both treatment assignment and outcome. Twenty-four sets of assumptions about the hypothetical covariate’s relationship to treatment assignment and outcome were examined; the most extreme of these tripled the odds of surgery and tripled the odds of improvement for a subset of patients. Under the various assumptions, the estimated proportion of patients experiencing functional improvement following medical therapy varied from 0.34 to 0.38, while those experiencing improvement after surgery varied from 0.63 to 0.70. In other words, none of the assumptions would have changed the original conclusions of this analysis.

<sup>27</sup>Specifically, “nonswitchers have 6 percent fewer prescriptions than switchers during the post-intervention period ... [used] less than ... one-sixth of the total dose ... [used] fewer other drugs and experience[d] fewer medical encounters .... [T]he estimates are several standard errors from zero” (Rubin, 1991, p. 1225).

Of course, the value of any particular sensitivity analysis depends upon a number of factors concerning the particular assumptions used. Most important is whether the varied assumptions specified by the analyst actually capture the true potential for results that are different than those that were observed.

A different type of sensitivity analysis has been suggested by Rindskopf (1986): use multiple alternative adjustment methods and check results for convergence and robustness.

## Goodness-of-fit Tests

Goodness-of-fit tests compare (1) hypothetical or predicted outcomes based on a model (or hypothesized distribution) to (2) observed outcomes. Goodness-of-fit tests can indicate how much difference the adjustments that the analyst used actually made—and thus signal the analyst to instances where adjustments make no difference or where “overadjustment” mars results.

For example, one goodness-of-fit test (used in Krakauer and Bailey, 1991, pp. 526-27) was based on the proportion of “concordant pairs.” That is, “in pairs consisting of a patient who did die and one who did not die,” the concordant pairs are those “pairs in which the patient who died had the higher probability of dying” according to the adjustment model. The higher the proportion of concordant pairs, the better the apparent fit of the adjustment model.

Specifically, Krakauer and Bailey used this goodness-of-fit test to assess the adequacy of risk adjustment achieved in an analysis of “variations in mortality rates among acute-care hospitals treating Medicare beneficiaries” (p. 527). Successive goodness-of-fit tests were conducted for each of four sets of adjustments considered: demographics only; demographics plus other claims information;<sup>28</sup> demographics and claims plus clinical findings; and all these plus information on the hospital (see table III of Krakauer and Bailey, 1991, p. 527).<sup>29</sup> The adequacy of adjustments based on demographics and other claims data was of special concern, because such adjustments are routinely used in published mortality rates for hospitals.

<sup>28</sup>Claims data include demographic information, the reason for the patient’s hospitalization, other chronic conditions, and prior hospitalizations.

<sup>29</sup>The concordant pairs test was successively applied when using the four different sets of adjustments (demographics only; demographics plus claims information; demographics, claims, and clinical; and all these plus hospital). The proportion of concordant pairs was 0.64, 0.84, 0.90, and 0.90, respectively. This assessment suggested that adjustments based on claims data were adequate; confirming tests were conducted based on rank correlations.



For further discussion of goodness-of-fit tests, model specification tests, and related topics, especially as these apply to structural equation modeling and to selection modeling, see Bentler and Bonett (1980) and Heckman and Hotz (1989a, 1989b). Of course, all such methods are limited by the set of measured covariates; hidden biases may remain undetected.

---

### Conducting Secondary Analyses to Test for Imbalance

When analyses such as those described above have not been performed by the data base analyst, the cross design investigator can sometimes conduct them. The advantages and difficulties of secondary analyses have been outlined by Boruch, Cordray, and Wortman (1981). Clearly, secondary analysis is possible only when the original data set is both available and adequately documented. Even then, difficulties may arise when attempting to match a specific subset of data or to recreate specific definitions and adjustments. The time required for secondary analysis may also present a barrier, especially if several different data bases are involved.

Where secondary analysis of data is feasible, however, the cross design investigator can perform the kinds of tests described above. In particular, secondary analysis provides the investigator with the opportunity to expand upon the no-difference tests reported by the primary analyst and to perform “tests of spuriousness.”<sup>30</sup>

### Expanding “No-difference” Tests Beyond Those Reported by the Data Base Analyst

Accessing the data base greatly expands the number—and type—of possible comparisons beyond the limited ones made by the original data base analyst. For example, among the criticisms of the usual comparisons of baseline characteristics seen in medical journals is that only the means (average values) of baseline characteristics in each group are reported, without the standard deviations of those characteristics (Altman and Doré, 1990). Also common is the comparison of baseline characteristics “one variable at a time” (Lavori et al., 1983, p. 1291). This is another practice that secondary analysis might improve upon.

A cross design investigator’s secondary analysis would ideally examine entire distributions for such variables as patient age at diagnosis

---

<sup>30</sup>Spuriousness is the term used by Lazarsfeld and by Rosenberg when (1) an alleged effect, such as A causes B, is based on an observed correlation of A and B, and (2) a prior factor C has caused both A and B, so there is no intrinsic link between A and B (see Lazarsfeld, Pasanella, and Rosenberg, 1972; Rosenberg, 1968).

separately for each comparison group.<sup>31</sup> A secondary analysis could also compare joint distributions on key variables—for example, age by a disease severity indicator—across comparison groups. Where possible, it is preferable to consider the overall or combined impact of “the difference in the distribution of the background variables” (Mosteller, 1990a). When multiple variables are involved, a propensity score approach (Rosenbaum and Rubin, 1984) could make such an analysis more feasible. Outcomes unrelated to the treatment’s effect could also be used in no-difference tests, as discussed above.

## Conducting Tests of Spuriousness

A secondary analysis can test for whether an observed effect is spurious. A test of spuriousness begins with subclassification (Cochran, 1965) on patient baseline characteristics. The treatment effect is calculated separately for each subclass (or stratum).<sup>32</sup> The pattern signaling spuriousness is that the treatment effect disappears (or is substantially reduced) within each subgroup. For example, a substantial observed effect for all patients would be spurious if within each age stratum the effect of the treatment is close to zero.<sup>33</sup>

A secondary analysis testing for spuriousness would use the data as adjusted by the primary analyst. The subclasses or strata would be formed using baseline characteristics that were either (1) not adjusted in the primary analysis or (2) possibly incorrectly adjusted (in the reviewer’s judgment).

## Comparing Results Across Studies

Even when the cross design investigator cannot conduct a secondary analysis, two other types of empirical assessments can be used to assess the impact of imbalances in comparison groups. The first compares treatment effects across data base analyses that differ in terms of the

<sup>31</sup>The reason is that even if two groups have identical means, important differences in distributions could occur—perhaps indicating that different portions of the patient population are represented in the comparison groups (e.g., that young and old patients are in one group, while middle-aged patients are in the other group).

<sup>32</sup>Subclassification has sometimes been termed “poststratification” (Lavori et al., 1983, p. 1294). Both terms refer to the process of dividing the data into patient groups or strata formed by one or more baseline characteristics (such as patient age).

<sup>33</sup>This pattern signals that the global all-subjects effect derived, at least in part, from different distributions of age in the comparison groups. The reason is simply that the age groups more likely to have a better-than-average outcome were concentrated in one of the comparison groups.

Although we have emphasized “subgroup” analysis in tests of spuriousness, this approach may also be carried out using techniques such as partial correlation or multiple regression to control for additional baseline characteristics. That is, one controls for the imbalanced factor and observes whether the partial correlation coefficient is substantially lower than the originally observed effect.



degree of known or likely imbalances in comparison groups. The other compares data base patients' outcomes following a specific treatment (such as their average blood pressure after taking a drug designed to reduce it) to outcomes for those patients in a randomized study who received the same treatment. Both types of assessment are described below.

### Comparing Effects Estimated by Different Data Base Analyses

To assess whether apparent imbalances in comparison groups actually impacted estimated treatment effects, meta-analysts have made cross-study comparisons. For example, Wortman and Bryant (1985) divided studies of school desegregation and academic achievement into two strata: (1) those with treatment assignment bias (known or likely imbalances) and (2) those without detected bias. The average effect size for studies deemed to have no bias was only 0.20, whereas for potentially biased studies, it was 0.50.

Comparison of effects across multiple data base analyses is really a two-part approach. The cross design investigator (1) draws upon the methodological and empirical assessments described earlier in the chapter to rate each data base analysis for comparison bias, and then (2) compares treatment effects across these strata. This shows whether data base analyses with apparent imbalances reported higher (or lower) effects than those without such imbalances; that is, it tests whether—and how—detected imbalances affected results.<sup>34</sup>

Of course, before comparing effects across data base analyses, the investigator should rule out the possibility that variability among findings derives purely from sampling error. Heterogeneity tests have been devised to address this problem (see Hedges and Olkin, 1985).

### Comparing Outcomes Following a Given Treatment

Analyses from the Duke Cardiovascular Disease Databank have compared (1) Duke patients' survival following a medical therapy to the survival of patients who received the same therapy in randomized studies, and (2) Duke patients' survival following surgery to the survival of patients who received the same surgery in randomized studies. The Duke data had previously been adjusted according to a model of the disease. Results showed that, for 24 out of 26 comparisons, the adjusted Duke estimate of

<sup>34</sup>This two-part approach follows a long-standing tradition: comparing effects observed in randomized and nonrandomized studies to test for treatment assignment bias in the nonrandomized study. (For a brief review of cross design comparisons of effects, which date back to work by R.A. Fisher in the 1930s, see Boruch, 1987, pp. 328 ff.) We believe such comparisons are best made after all assessments are completed and after secondary adjustments of individual study results have been made by the cross design investigator (see chapter 4).

patient survival was within the 95-percent confidence limit surrounding the randomized study estimate. This demonstrated the absence of comparison bias in the adjusted Duke data.<sup>35</sup>

The details of this important study are as follows: Three sets of patients were selected from the Duke Cardiovascular Disease Databank to match the entry criteria of three randomized studies. Outcomes for each set of Duke patients were adjusted using a statistical model. Finally, the adjusted 5-year survival rates for the Duke patients receiving each therapy were compared to those for patients receiving the same therapy in a randomized trial (Hlatky et al., 1988; see also, Pryor et al., 1983; Califf et al., 1986).

The adjusted patient outcomes from the databank matched patient outcomes reported in three published randomized studies, showing that balance had been achieved in the Duke data base analysis.<sup>36</sup> Given that the Duke Databank example compared outcomes to show an absence of comparison bias, the next question is: Can the same type of assessment be used to show the presence of comparison bias? We believe such assessments are possible.

Briefly, the logic of using cross design comparisons of outcomes to check for signals of comparison bias rests on a simple fact: patients' outcomes reflect—in part—their prognoses prior to treatment. Other factors also contribute to patient outcome levels and must be taken into account. Notably, when outcomes of randomized studies are part of the pattern, the reviewer must be alert to the possibility that the patient pool (consisting of treatment and control groups combined) may not be representative. That is, patients participating in randomized studies may have had initial prognoses that differed from those of the full population of patients. Nevertheless, the expectation is that, within a randomized study, the prognoses of the treatment and control groups are alike. With this in mind—and ideally assuming that some notion of the level of nonrepresentativeness in each randomized study was gleaned through assessments described in chapter 2—the cross design investigator may sort out appropriate interpretations for various possible patterns.

<sup>35</sup>This assessment of comparison bias applies only to data base analyses using a treatment group comparison. It cannot be used to diagnose imbalance in a natural cohort comparison.

<sup>36</sup>A different type of analysis has been performed comparing effects estimated in the Duke data set to effects estimated in randomized studies (see Hlatky, 1991); such comparisons are discussed in chapter 4 of this report. Note, however, that an effect represents the difference in survival that a treatment made (as opposed to survival itself). A comparison of outcomes (e.g., a comparison of survival rates) following each treatment provides more information than a comparison of treatment effects. This additional information is needed for the reviewer to diagnose the presence of comparison bias as opposed to other problems.



Table 3.2 is presented as a preliminary guide to interpreting patterns of patient outcomes in data base analyses and randomized studies. That is, table 3.2 distinguishes patterns that signal comparison bias in a data base analysis from patterns that point to other kinds of problems.<sup>37</sup>

**Table 3.2: Logic of Patient Outcome Comparisons: Patterns Signaling Comparison Bias in a Data Base Analysis**

Outcomes for patients receiving usual care	Outcomes for patients receiving new treatment		
	$\bar{X}_t > \hat{\mu}_t$	$\bar{X}_t < \hat{\mu}_t$	$\bar{X}_t = \hat{\mu}_t$
$\bar{X}_{uc} > \hat{\mu}_{uc}$	Patients in randomized study may not be representative	<b>Comparison bias likely in data base analysis</b>	Combination of problems
$\bar{X}_{uc} < \hat{\mu}_{uc}$	<b>Comparison bias likely in data base analysis</b>	Patients in randomized study may not be representative	Combination of problems
$\bar{X}_{uc} = \hat{\mu}_{uc}$	Treatment implementation may differ: data base versus randomized study	Treatment implementation may differ: data base versus randomized study	Convergence

$\bar{X}_t$  = mean outcome for the treatment group of a randomized study

$\bar{X}_{uc}$  = mean outcome for the control group of a randomized study (in which patients were assigned to "usual care")

$\hat{\mu}_t$  = mean outcome for data base patients who received the new treatment

$\hat{\mu}_{uc}$  = mean outcome for data base patients who received usual care

Note: This table is a guide to distinguishing comparison bias in data base analyses from two other problems: (1) nonrepresentativeness of patients in randomized studies and (2) treatment implementation differences between randomized studies and data bases. The logic of the table assumes that any other potential problems were ruled out.

Table 3.2 is essentially the same as table 2.1. The only difference is that table 3.2 highlights patterns signaling comparison bias, whereas table 2.1 highlights patterns signaling lack of generalizability.

The patterns described in table 3.2 are limited to mean (average) outcome levels. However, patterns in the standard deviation of the outcome variable could also be examined across studies. Generally, in the absence of comparison bias, one would expect patterns such as the following:

<sup>37</sup>Caveats for interpreting table 3.2 are noted at the bottom of the table. However, we believe that this approach—although preliminary—is a substantial improvement upon the more traditional practice of comparing effects estimated in data base analyses to those observed in randomized studies and concluding that treatment assignment bias in the data base estimate is likely whenever there is a difference in effects.

- For patients receiving the new treatment, the standard deviation for data base patients should be as large as—or larger than—the standard deviation for patients in a randomized study.
- For patients receiving usual care, the standard deviation for data base patients should be as large as—or larger than—the standard deviation in a randomized study.
- The ratio of the standard deviations for two groups of data base patients receiving alternative treatments should be the same as the ratio of the standard deviations for treatment and control groups of randomized studies.

## Summary of Task 2: Steps in Assessing Comparison Bias

Some weakness is associated with each method of assessing comparison bias. But when the methods are jointly applied to the same data base analysis (or to each data base analysis in a set), the combined results should either (1) indicate the balance or imbalance of the patient groups being compared, or at least (2) point to a persistent uncertainty about the nature and degree of comparison bias.

Five specific steps that a reviewer would follow to assess imbalanced comparison in each existing data base analysis are listed in table 3.3. These steps are based on the foregoing discussion of assessments that have been aimed at the two major factors involved in imbalanced comparisons:

- the methods of comparison and adjustment used in the data base analysis (steps 1 and 2); and
- the achieved balance or imbalance of the adjusted comparison groups (steps 3, 4, and 5).

Steps 1 and 2 require judgmental assessments. Steps 3 through 5 require empirical analyses.



**Chapter 3**  
**Methods for Assessing Data Base Analyses**  
**for Comparison Bias (Task 2)**

**Table 3.3: Assessing Imbalanced Comparisons In Data Base Analyses: Five Steps**

Target of assessment	Major steps	Assessment methods for conducting each step
Methods of comparison and adjustment	<b>Step 1.</b> Assess data base analyst's choice of treatment groups versus "natural cohorts."	Judge whether treatment groups are independent of patient prognoses; repeat for natural cohorts (Moffitt, 1991; see table 3.1 of this report).
	<b>Step 2.</b> Assess adjustments used to balance comparison groups.	Judge completeness of covariates adjusted for by, e.g., checking models of the disease (Pryor et al., 1983), generic case-mix indicators (e.g., Wagner et al., 1983; Charlson et al., 1987).  If possible, check potential for invalid adjustment (e.g., very large adjustment needed or measurement error in covariate).
Achieved balance in comparison groups	<b>Step 3.</b> Review empirical tests provided by the data base analyst.	Review data base analyst's tests for:  —"no difference" in comparison groups (see Yeaton, 1990; Rubin, 1991);  —sensitivity of the adjusted results to an unmeasured covariate (see Rosenbaum and Rubin, 1983b); and  —the adjustment model's goodness of fit (Kraakauer and Bailey, 1991; Bentler and Bonett, 1980).
	<b>Step 4.</b> Perform secondary analyses (if feasible).	Perform tests on adjusted data to:  —supplement empirical tests performed by the data base analyst (see step 3) and  —check for "spuriousness" (Lazarsfeld, Pasanella, and Rosenberg, 1972).
	<b>Step 5.</b> Compare data base results to other studies' results.	Compare:  —the treatment effect to effects observed in other data base analyses (Wortman and Bryant, 1985; Lipsey, 1992). <sup>a</sup>  —patient outcomes in the data base analysis to patient outcomes in randomized studies (see Hlatky et al., 1988; table 3.2 of this report).

<sup>a</sup>Check whether analyses judged likely to be imbalanced in steps 1 and 2 yielded larger or smaller effects than more balanced analyses.

Once steps 1 through 5 have been completed for an existing data base analysis, the cross design investigator combines this information to form a judgment about the nature and extent of imbalance associated with the results of that data base analysis. Alternatively, the investigator may conclude that uncertainty remains, perhaps because there was a lack of information needed to complete the assessments.

As was noted in the previous chapter, these assessment steps have not been previously presented as a set of analyses. Thus, there are a number of potential technical issues that will need to be raised and resolved concerning their use in combination.



# Methods for Adjusting and Combining Results of Randomized Studies and Data Base Analyses (Tasks 3 and 4)

Once the assessments discussed in the previous chapters have been conducted, the cross design investigator will likely have accumulated considerable information on the weaknesses in the studies to be combined. Specifically, task 1 of the cross design synthesis should have indicated the nature and extent of generalizability problems for existing randomized studies. Task 2 should have revealed much about comparison bias in the results of each data base analysis. Alternatively, the investigator will have become aware of the uncertainty associated with existing results. Assuming that a cross design synthesis has been deemed both needed and feasible, the investigator now faces the major challenges of this approach.

As noted at the outset of this report, even though the different designs were selected because they have complementary strengths and weaknesses, one cannot naively combine results across categories, trusting that the weaknesses will “average out,” while the complementary strengths are preserved. Rather, the cross design investigator must successfully complete the two remaining tasks:

- Task 3 is to perform secondary adjustments, as needed, to minimize problems and biases in existing studies.
- Task 4 is to synthesize the adjusted results within and across design categories, recognizing the limitations of these results and taking account of persistent cross-study differences.

In this chapter, we further describe the challenges and logic of tasks 3 and 4, the steps involved, and the methodological options available to the cross design investigator.

## Logic of Task 3: Adjusting Individual Studies’ Results

The cross design investigator must take account of known biases in individual study results.<sup>1</sup> Two alternatives for dealing with known biases in individual studies have been put forward previously. These are:

- exclusion of the more biased studies, and
- secondary adjustment of individual study results to compensate for specific biases.

Exclusion (that is, assigning the more biased studies a “zero weight”) is not advocated here as a solution to the challenge of known biases in

<sup>1</sup>Based on tasks 1 and 2, the investigator will have information on what portions of the patient population are inadequately represented in each randomized study and on the direction of comparison bias in each data base analysis.

existing studies. Of course, some studies may in fact be so irretrievably biased or flawed that a “zero weight” is necessary. But in a cross design synthesis, such an approach does not represent an ideal first solution to the problem of individual study bias.

The reason is that although each study is weak in an important area, it has an alternative strength. Our goal is to capture these strengths. For example, a randomized study is likely not fully generalizable to the patient population, but it should be included in a cross design synthesis because of its strength in providing a valid comparison. The situation is analogous for data base analyses. Exclusion signifies a failure to reap the benefits of the strengths contained within a particular study. Thus, exclusion on the basis of a known bias should represent a last resort.

A more promising approach is to begin with the secondary adjustment of each study’s results to counteract the specific form of bias that threatens their validity. Eddy’s “confidence profile method” strongly advocates adjusting the results of each study before developing a combined estimate (Eddy, Hasselblad, and Shacter, 1989, p. 56):<sup>2</sup>

“the assumption being made is that all studies to be combined... [in order to estimate]...a particular parameter must be estimating that parameter without bias, or must have been adjusted for any biases that affect their estimates of that parameter.”

More specifically, Rubin (1990a) has suggested that before including randomized studies in a meta-analysis, their results should be adjusted to reflect population distributions (of patients’ characteristics, for example). This produces meta-analysis results that go beyond mere representation of the kinds of subjects that happen to have participated in the various studies. Mosteller (1990b, citing Colditz, Miller, and Mosteller, 1988) endorses secondary adjustment for comparison bias in results of nonrandomized studies. As Mosteller notes, however, practicing meta-analysts have only infrequently applied secondary adjustments.<sup>3</sup>

---

<sup>2</sup>The confidence profile method is a strategy for statistically combining multiple pieces of evidence from different experimental designs involving different types of outcomes, different measures of effect, and different kinds of biases. This method also utilizes indirect evidence (on, e.g., intermediate outcomes) and “mixed comparisons” (“one experiment might compare treatment A with treatment B, another treatment B with treatment C”). It advocates using nonexperimental, subjective evidence where there are gaps in formal knowledge (see “Executive Summary” of Eddy, Hasselblad, and Schacter, 1989).

<sup>3</sup>A considerable literature exists on primary adjustments. Indeed, epidemiologists point out that adjustments should be made whether or not an observed bias reaches statistical significance: “Even if such a distortion occurs by chance...it would still have to be corrected in the data in order to obtain a proper estimate....” (Kleinbaum, Kupper, and Morgenstern, 1982, p. 254).



One precedent for individualized secondary adjustments is provided by a GAO report (GAO, 1989b) concerning AIDS forecasts. GAO first assessed the different undercount and overcount problems in the AIDS data base, which was used by all the forecasters. GAO then determined which problems had (and which had not) already been adjusted for in each study. This allowed secondary adjustments to be individually tailored to each existing forecast.

In a cross design synthesis, individualized adjustments that take account of the estimated distortion in each study's results are possible because of the in-depth assessments conducted in tasks 1 and 2. Thus, we advocate individualized secondary adjustments to account for known biases. The first step consists of the secondary adjustment of randomized studies' results to enhance generalizability; that is, standardizing results to the relevant patient population. The second step consists of adjustment of data base analyses to minimize comparison bias through a variety of methodological options. Methods for completing these steps are described later in this chapter.

---

## Logic of Task 4: Combining Results Within and Across Design Categories

An important challenge facing the cross design investigator is the possibility that, despite secondary adjustments, the chief weaknesses associated with the major designs continue to distort study results. Notably:

- Some patient groups may have been totally excluded from randomized studies. This is a problem that cannot be fixed by adjustments of individual studies' results to correct for overrepresentation and underrepresentation (in task 3).
- In the assessment of each data base analysis, some imbalances in patient comparison groups may not have been detected. If any imbalance is not detected by either the data base analyst or the cross design investigator (and thus was not adjusted in task 3), hidden comparison bias will remain in the adjusted data base results.

One solution is to devise an appropriate framework for organizing, analyzing, and combining results from different categories of study designs. This approach has its roots in meta-analysis, where strata based on study designs or characteristics of study participants (Light and Pillemer, 1984) have been created so that results from multiple, relatively homogeneous studies could be analyzed separately and combined within each stratum. Such strata might be used to "take into account different

characteristics of subjects, treatments, contextual variables, and effects of interactions among these” (Jackson, 1980, p. 135).

Thus, we advocate that, following secondary adjustment of individual study results, the cross design investigator devise a synthesis framework. The purpose of this framework is to allow differences in design—and potentially persistent biases—to be accounted for when combining studies within and across design categories.

Combining multiple studies within a major design category (or stratum) involves the traditional challenges of meta-analysis. For example, the results of randomized studies may differ from each other—even after secondary adjustment of each study’s results for known artifacts. Such differences should be analyzed and taken into consideration when deciding how the results of randomized studies are to be combined. The same logic applies when combining results from multiple data base analyses.

Even after adjustment, differences in results of individual studies may remain and be linked to differences in:

- the certainty that the investigator has about different studies’ results, deriving from a number of sources, including (1) known differences in the quality of the studies and (2) differences in the investigator’s knowledge about the nature and extent of biases remaining in the studies’ adjusted results;
- study procedures or design specifications (for example, differences in the target populations of various randomized studies, such as male patients only in one study and female patients only in another study); and,
- the reliability of different studies’ results (deriving from different sample sizes, variances).<sup>4</sup>

Differences in certainty should not be ignored, as most meta-analysts would agree (although specific approaches for dealing with differences in certainty have varied). Likewise, differences in procedures or design specifications should be recognized and taken into account via a logical model, stratification, or other method. Many analysts have realized this in recent years. And the more reliable studies should be given greater weight, as statisticians have shown (see Rao, 1984, citing Cochran, 1937, 1954, and Cochran and Carroll, 1953). Thus, we believe that when combining studies

---

<sup>4</sup>This limited list of possible sources of differences in adjusted results reflects our continuing assumptions in this report: a constant treatment implementation and a single outcome of interest.



within each major design category, the investigator should take account of cross-study differences through such methodological options as appropriate weights or the use of ranges, which are described later in this chapter.

After a multistudy estimate of the treatment effect has been calculated separately for each major design category, the cross design investigator faces what may seem to be his or her greatest challenge. In order to reap the benefits of the cross design approach, the investigator must successfully deal with two potential problems:

- First, one or more logical design categories may be “empty sets” in the framework; notably, there may well be no existing randomized study that covers a major patient population group (such as elderly patients).
- Second, estimates of the treatment’s effect may differ across design categories, and differences in these estimates may reflect (or derive from) differences in the levels of certainty associated with the various estimates, differences in study design and population coverage, and of course, differences in the reliability of these estimates.

In combining studies across design categories, the investigator must project results to the empty stratum and must determine the appropriateness of combining estimates of treatment effects across other design categories. This involves considerable investigator judgment about how (or if) to combine results from specific strata. For example, an investigator may decide to take a weighted average of results across certain design categories. Or, estimates from different design categories might be used to define ranges for plausible sizes of the treatment’s effect.

## Tasks, Steps, and Challenges in Adjusting and Combining Studies

The task of adjusting individual studies’ results involves two major steps: (1) secondary adjustment of each randomized study’s results to correct for known overrepresentation and underrepresentation of patient groups (that is, standardization to population distributions); and (2) adjustment of the results from each data base analysis to compensate for known comparison bias. The task of combining studies’ results involves three additional steps: (3) constructing a framework for separately analyzing results of studies in different design categories; (4) combining studies within each design category, while taking account of differences between studies that may have affected results; and (5) synthesizing results across design categories, again taking account of major differences in design that may have affected results.

These steps have been designed to meet the challenges involved in adjusting and combining studies, as summarized in table 4.1. In the sections that follow, we review methodological options that the cross design investigator can draw upon to complete these steps.

**Table 4.1: Adjusting and Combining Studies: Challenges, Tasks, and Steps**

Challenge	Task or step
General challenge: individual study results are known to be biased by artifacts identified in assessment.	Task: Adjust each study's observed treatment effect before combining.
<b>Specific challenge:</b> randomized study results are not generalizable.	— <b>Step 1.</b> Standardize each randomized study's results to patient population parameters.
<b>Specific challenge:</b> data base comparisons are imbalanced.	— <b>Step 2.</b> Lower or raise effects estimated in data base analyses, using assessment results.
General challenge: Studies' results differ both within and across major categories of study design and population coverage.	Task: Combine adjusted study results, taking account of persistent differences, both within and across design categories.
<b>Specific challenge:</b> different strategies must be used to take account of differences within and across major design categories.	— <b>Step 3.</b> Create synthesis framework (design strata) to organize results.
<b>Specific challenge:</b> within each design category, studies' adjusted results may differ according to their procedures, reliability, and quality.	— <b>Step 4.</b> Within each stratum of the synthesis framework, combine results using a plan to adjust for multiple types of cross-study differences.
<b>Specific challenges:</b> there is a lack of results for one or more key categories (e.g., a lack of randomized study results for important patient groups); multistudy estimates do not converge across design categories.	— <b>Step 5.</b> Use synthesis framework to develop cross design projection; make judgments about using "better" estimates (only) versus ranges or weighted averages taken across strata

## Making Secondary Adjustments to Enhance Generalizability

The purpose of adjusting each randomized study's results is to enhance generalizability; that is, to correct for the overrepresentation and underrepresentation of certain patient groups. As Rubin (1990a) has suggested, a randomized study's results can be weighted to reflect known parameters of the patient population (such as patient age distribution). The cross design investigator can accomplish this without secondary data analysis—provided that separate treatment effects were reported for relevant patient subgroups (for example, a separate treatment effect for each age group).

Specifically, the cross design investigator applies a set of weighting procedures that have been in common use for many years in survey sampling and in epidemiological research. In survey sampling, the goal has



been to improve the match between sample survey results and the corresponding values in the population from which the sample was drawn. The sample-survey weighting procedures include:

- “adjusting sample frequencies to expected marginal totals” (Deming, 1964, p. 96);
- the “nights-at-home” adjustment (Politz and Simmons, 1949) used in many opinion polls;<sup>5</sup>
- weighting random (second-wave) subsamples of first-wave nonrespondents to represent all first-wave nonrespondents (Cochran, 1963, citing Hansen and Hurwitz, 1946);
- weighting survey respondents in a given subgroup or geographic area by, for example, the inverse of the response rate for that subgroup or geographic area (see Kalton, 1983); and
- weighting survey respondents in various subgroups or geographic areas so that the resulting distributions across these subgroups or areas will match published data from the U.S. Census (Kalton, 1983).

In epidemiologic research, essentially the same approach has been used to produce a standardized incidence (or prevalence) rate. Here, the classic situation consists of two populations that differ on a characteristic (such as age) that affects the incidence rate of the disease in question. To better compare incidence rates across the two populations, these rates have been standardized for one or more key characteristics. For example, a drug use incidence rate in the armed forces and in a civilian population might be standardized for age and sex.

When only one key characteristic is taken into account, the standardization procedure is straightforward. First, a standard population is identified and information on the distribution of the key characteristic (such as age) is obtained. Then, as described in Liberati et al. (1988), disease incidence rates for members of a study sample are calculated separately for each category of the key characteristic (for example, for each age category). Lastly, a weighted average of the category-specific disease incidence rates is calculated, with weights taken from the (age) distribution in the standard population. Deming (1964) and Fleiss (1973)

<sup>5</sup>To illustrate the way the “nights-at-home” adjustment works, a relative weight of 4 is given to each survey respondent who reports that he or she was at home on only one of the four preceding nights; each such respondent thus represents himself or herself and three other (presumably similar) persons who were not interviewed because they were not at home that night. By contrast, a survey respondent who reported being at home all four nights would receive a weight of 1, standing for himself or herself only.

describe procedures for calculating rates standardized on multiple characteristics.

Fortunately, because these methods are based on weighting grouped data, they lend themselves to secondary adjustment of published results (provided, of course, that the required subgroup data have been reported). For example, suppose that a randomized study includes equal numbers of younger patients (31-50 years) and older patients (51-70), but that only 20 percent of the total patient population falls into the 31-50 group. The large majority of the patient population (80 percent), then, are in the 51-70 age group. If the randomized study reports separate estimates of the treatment effect (for example, reduction in high blood pressure) for patients aged 31-50 years and for patients aged 51-70, then it is possible to calculate an overall treatment effect that is “standardized” or adjusted to reflect the age distribution of the patient population—at least within the 31-70 age range. The procedure is to take a weighted average of the separate treatment effects, using the information on the population age distribution as weights; for example, 0.20 times the reported effect for the younger group plus 0.80 times the reported effect for the older group. Such an adjustment would counteract overrepresentation and underrepresentation in each of these age groups.

---

## Making Secondary Adjustments to Minimize Comparison Bias

The traditional approach to minimizing comparison bias in existing studies is the secondary analysis of the data base in question. This means accessing the data base and using methods of primary adjustment to control for “confounding variables.” Methods of primary adjustment include matching, subclassification, and adjustments based on weighted averages taken across strata, partial correlation, covariance or regression adjustment, propensity scores, selection modeling, or structural modeling (see chapter 3).

Although secondary analysis appears to be the safest approach, it is not always possible because of nonavailability of the original data set or other problems. In addition, the time and resource requirements for conducting secondary analyses of multiple data bases may be prohibitive. When secondary analyses are not possible, the cross design investigator may turn to other options, such as:

- secondary adjustment using a variant of the standardization procedures discussed above (that is, a variant of the procedures recommended for standardizing a randomized study’s results);



- secondary adjustment of the observed treatment effect, upward or downward, using a ratio (to, for example, reduce the observed effect by 10 percent); and
- calculation of an alternative estimate of the treatment effect based on “natural experiments” or natural cohort comparisons (as discussed in chapter 3).<sup>6</sup>

A “standardized treatment effect” that corrects for comparison bias can be calculated by taking a weighted average of treatment effects for each separate subgroup defined by the key characteristic or prognostic factor. For example, the separate treatment effects observed for each age group could be averaged—using the relative sizes of the age groups as weights.<sup>7</sup> However, if subgroup effects were reported by the data base analyst, it is likely that related differences were taken into account in the primary adjustments.

Adjustment of the observed treatment effect upward or downward by a ratio has been described by Eddy, Hasselblad, and Schacter (1989, pp. 107-08). The investigator specifies “a ratio for the outcome parameter that applies to individuals in the treated group compared with individuals in the control group, in the absence of the intervention.”<sup>8</sup> The benefit of using a ratio adjustment is that this avoids the need for accessing the data base. The important issue, however, concerns the basis used to set the ratio. The in-depth information gleaned in the assessment of comparison bias (chapter 3) would inform the investigator’s subjective choice of a ratio for each individual study.<sup>9</sup>

---

<sup>6</sup>The options for secondary adjustment of data base results discussed in this section do not include use of randomized study results as a standard. In the strategy of cross design synthesis, comparisons of treatment effects estimated in randomized studies and in data base analyses are not made until the last step. The reason is that comparisons across design categories are most appropriately made (1) after studies have been individually adjusted based on in-depth assessment of their specific weaknesses and (2) after the best multistudy estimate has been derived for each major design category.

<sup>7</sup>It is also possible to take a weighted average of outcomes for the treatment group, and then for the control group; that is, to standardize the groups separately. The weighted results for the treatment group could then be compared to the weighted results for the control group. Of course, this approach can only be used if outcomes were reported for separate age groups of patients—within each comparison (or treatment) group.

<sup>8</sup>For example, suppose that the patients who received the new treatment would have been—in the absence of that treatment—twice as likely to experience a negative outcome as patients who received usual care. Logically, then, the usual care group should be subjected to a compensatory adjustment—doubling the proportion of usual care patients experiencing the negative outcome. The investigator would make this secondary adjustment before recalculating the treatment effect.

<sup>9</sup>Obviously, the ratio that is appropriate for adjusting one data base analysis might not be appropriate for adjusting another.

Calculating the difference between average outcomes for two similar populations of patients, each of which received a different treatment, constitutes taking advantage of a “natural experiment.” The work of Wennberg (for example, Wennberg et al., 1988) suggests that such possibilities do occur.<sup>10</sup> This approach follows the natural cohort comparison discussed in chapter 3. Published outcomes for such natural cohorts may be available; if so, an alternative estimate could be derived.

The use of different adjustment techniques has sometimes been found to yield different results. Therefore, a careful cross design investigator may wish to test the robustness of secondary adjustments by applying multiple, alternative techniques of secondary adjustment (see, for example, Rindskopf, 1986). Where necessary, a range of adjusted values can then be developed for each study; where such ranges are large, the uncertainty of the adjusted value is clear.

As deemed appropriate by the investigator, the estimates obtained via one or more of the foregoing adjustment options can be used either (1) as a substitute for the original data base estimate of the treatment’s effect, or (2) together with the original estimate as, for example, in a range.

## Designing a Synthesis Framework

Meta-analysts have often used separate design categories (or strata) to analyze study results. For example, the treatment effects observed in randomized studies and “quasi-experiments” were analyzed separately in Wortman and Yeaton’s (1983) meta-analysis of studies of coronary artery bypass graft surgery. And results for studies covering premenopausal and postmenopausal breast cancer patients were analyzed separately in a meta-analysis of randomized studies of adjuvant chemotherapy (Himel et al., 1986). Following separate analysis, the studies in the different categories may or may not be combined in a single estimate.

Outside the meta-analysis tradition, one notable study (Hlatky, 1991) suggests the value of a framework that (1) stratifies observed treatment effects on both study design and population coverage and (2) fine-tunes the population coverage strata (so that the results for randomized studies and data base analyses can be compared for matched patient groups).

<sup>10</sup>This work has revealed surprisingly high levels of variation in medical practice—including differences in the preferred treatments for specific diseases and conditions—across geographic areas. As a result, patients in certain geographic areas where the new treatment is very widespread may be overall very similar to patients in other geographic areas where that treatment is rare.



As discussed in chapter 3, researchers at Duke compared outcome levels from three previously published randomized studies to results of an analysis of the Duke Cardiovascular Disease Databank. In that analysis, patients from the Duke Databank were selected to match the patient pools of the randomized studies.<sup>11</sup> The most recent in a series of studies that emerged from this effort (Hlatky, 1991) compares the treatment effect observed in each randomized study to the corresponding treatment effect estimated from records of comparable patients in the Duke Databank (see figure 1 of Hlatky, 1991). This Duke study suggests that, when secondary analysis of the data base in question is possible, design categories or strata should be based on “fine-tuned” definitions.

Clearly, any number of strata might be used to create a framework for synthesizing results in specific areas. A cross design synthesis to estimate a treatment’s effect across the full range of patients should include stratification on at least two key dimensions:

- The primary dimension of stratification corresponds to the types of designs included in the synthesis. In the present instance, this means that randomized studies form one stratum and data base analyses form another.<sup>12</sup> All further stratification is carried out within each primary design stratum.
- The secondary dimension of stratification is defined according to coverage of the patient population by existing randomized studies. The key strata here are: (1) subgroups of the patient population covered by the existing, combined set of randomized studies (even if the subgroups are underrepresented or overrepresented), and (2) subgroups of the patient population not covered at all in any existing randomized studies.

Table 4.2 depicts four strata for a simple version of the framework suggested here.

---

<sup>11</sup>Specifically, the patients were chosen from the Duke Databank to match eligibility criteria (including age, sex, type of heart disease, and patient entry date) for each of the three randomized studies (Hlatky et al., 1988).

<sup>12</sup>Of course, each dimension of stratification could include several strata (not just two). And more than two dimensions of stratification are possible. Thus, for example, when creating strata to take account of methodological variations, one might use several design strata instead of merely distinguishing between randomized studies and observational data bases.

**Table 4.2: Synthesis Framework:  
Primary and Secondary Dimensions of  
Stratification**

Secondary dimension: coverage of patient groups in randomized studies <sup>a</sup>	Primary dimension: type of design	
	Results of randomized studies: Stratum 1	Data base analyses: Stratum 2
Covered in randomized studies (e.g., whites)	Stratum 1a	Stratum 2a
Not covered in randomized studies (e.g., blacks, and other minorities)	Stratum 1b (empty)	Stratum 2b

<sup>a</sup>Assumes that existing data base analyses cover all patient groups.

To illustrate the secondary dimension of stratification, if an existing set of randomized studies covers only white patients—whereas approximately one-fourth of the patient population is composed of blacks and other minorities—then the secondary dimension of stratification subdivides the randomized study stratum and the data base stratum into racial strata as follows: randomized study results for white patients, randomized study results for nonwhite patients (an empty set), data base results for white patients, and data base results for nonwhite patients.

There are two reasons why the second dimension of stratification is necessary. The first reason is simply that certainty concerning randomized study results is quite different for patient groups covered by randomized studies and patient groups not represented in any randomized study. Second, the procedures that are appropriate for developing estimates of treatment effects are different for each of these strata, as described in the following section.

To obtain a very close match between the patient populations covered in stratum 1a and in stratum 2a, secondary analysis of the data base may be required, as evidenced by the Duke analysis described above.<sup>13</sup>

## Combining Study Results Within Design Categories

Having created a framework, the investigator's next step is to combine results from studies within each category (stratum), taking account of cross-study differences. As previously discussed, cross-study differences in estimated treatment effects may be linked to: (1) differences in the investigator's certainty concerning the results of the studies; (2) differences in the procedures and design specifications (such as

<sup>13</sup>The Duke analysis provided the pattern for strata 1a and 2a in table 4.2; however, it did not cover strata 1b and 2b.



population coverage) of the individual studies; and (3) differences in the reliability of the results of the various individual studies. To take account of each type of cross-study difference, specific methods must be identified and chosen; then, where multiple types of cross-study differences exist, a combined plan for taking account of them is needed. Available options are discussed below.

---

## Accounting for Differences in Certainty

There are at least three options for dealing with differences in certainty. These are: quality weights, projection (by, for example, extrapolation), and stratification. Many analysts advocate such approaches.<sup>14</sup>

The first option, long discussed by meta-analysts, consists of taking a weighted average of study results, weighting each study according to its quality. As Rosenthal (1984, pp. 54-55) observed:

“Once all the retrievable studies have been found, decisions can be made about the use of any study. Precisely the same decision must be made about any study retrieved: How shall this study be weighted? Dropping a study is simply assigning a weight of zero. If there is a dimension of quality of study (e.g., internal validity, external validity, and so on) then there can be a corresponding system of weighting. If we think a study is twice as good as another, we can weight it twice as heavily or four times more heavily and so forth.”

Rubin (1990a) has suggested another option: rather than averaging the results of “the current collection of fallible studies,” investigators develop ways to extrapolate to an “ideal study’s” results. Certainly, one can imagine a display in which observed treatment effects diminish with increasing study quality; if the size of the treatment effect appears to asymptote at a certain level of study quality, one would be reasonably confident in taking the effect size observed at that level as the “best estimate.” Alternatively, imagine a display in which the effect size decreases with study quality but does not appear to asymptote; one option would be to project an even lower value for the “ideal study” than any of those observed.

A third option is to create strata based on study quality (Jackson, 1980). This approach was used in the recent GAO report of forecasts of the AIDS epidemic (GAO, 1989b): Results were first presented for all forecasts identified (first stratum or category); they were subsequently presented for only those forecasts that met certain criteria (the highest quality stratum).

---

<sup>14</sup>Some, however, have argued against the use of the quality weights option (see, e.g., Eddy et al., 1989). It is true that if secondary adjustments improved all studies’ estimates to the point where differences in study quality were eliminated, such weights would not be relevant.

## Recognizing Differences in Study Specifications

There are at least three options for taking account of differences in specifications across studies of the same basic design (that is, options for taking account of differences such as one randomized study covers males, another covers females). They are: the use of an appropriate model, stratification, and ranges.

Rubin (1990a) has suggested that rather than averaging studies whose subjects represent different portions of a target population, the meta-analyst can extrapolate various studies' results to achieve an effect size for a defined target population.<sup>15</sup> Eddy, Hasselblad, and Schacter (1989) have suggested the use of explicit models (sketched out in "influence diagrams") to relate findings about effects, outcome levels, intermediate outcomes, and so forth—observed for various patient groups and influenced by specified cofactors.

The Rubin-Eddy cross-study approaches can be applied in ways that are very similar to the standardization of results for individual randomized studies (which was discussed in a previous section of this chapter). To suggest a simple example, suppose that one existing randomized study covers male patients only, while another covers female patients only, and the gender distribution of the patient population is 80-percent male, 20-percent female. A model might specify taking a weighted average with the weight of 0.80 assigned to the men's study result and the weight of 0.20 assigned to the women's study result.

Using the approach suggested by Rubin and by Eddy, the cross design investigator would achieve a single multistudy estimate for the design stratum in question (for example, a single estimate for all randomized studies, which is stratum 1a in table 4.2). However, in some cases, a range or set of estimates for a particular stratum might be deemed more appropriate (see, for example, GAO, 1989b). Alternatively, a range might be seen as a desirable addition to the best multistudy estimate for that stratum. Yet another option is suggested by Light and Pillemer (1984): When different types of studies or different types of subjects yield different results, one can stratify results by study characteristics, presenting them separately for each stratum, as appropriate. In the present instance, this would mean creating substrata within one or more major strata.

<sup>15</sup>In the previous section, we noted Rubin's projection approach as a way of taking account of differences in study quality; we note it here as a way of taking account of differences in population coverage.



---

Recognizing Differences in  
Reliability

Hedges (1984) demonstrates that the most precise multistudy estimate would be a weighted average, using the inverse of the variance of each study's estimate as that study's weight. In the same vein, Hunter and Schmidt (1990) advocate weighting by each study's sample size. Rosenthal (1984), whose meta-analysis method is based on combining z scores, similarly discusses the possibility of weighting z scores by the associated degrees of freedom.

It is important to note, however, that weighting study results to take account of reliability differences assumes that only estimates of a single treatment effect (that is, estimates of the same population value) are being combined. Thus, for example, when combining estimates of a treatment's effect on male patients with estimates of that treatment's very different effect on female patients, it would not be appropriate to weight one or the other category based on the sizes of the studies involved.

---

Dealing With Multiple  
Types of Study Differences

The various types of cross-study differences are not mutually exclusive. For example, the investigator may be faced with a set of randomized studies that differ from each other in terms of their quality (or the investigator's certainty about their results), their specifications, and their size (and hence their reliability). Fortunately, many of the approaches outlined above can be used in conjunction, to take account of multiple types of cross-study differences. This requires that a careful plan be devised by the investigator.

Suppose for example, that we are dealing only with studies within a single design category—stratum 1a in table 4.2. Further suppose that four randomized studies exist, all covering white patients only, and each has been standardized to reflect the age distribution of the patient population. Within this stratum or design category, the cross-study differences facing the investigator are as follows:

- One of the randomized studies is associated with a lower level of certainty than the others (that is, it has been judged “lower quality”).
- Three of the randomized studies cover male patients only and one covers female patients only (although the patient population is 50-50 male-female). Further, the one low-quality randomized study is a male-only study.
- All the randomized studies have very different sample sizes.

One plan for joint consideration of these various differences is shown in table 4.3. This plan produces two estimated treatment effects (estimate A and estimate B) for the total patient population. These might be viewed as a range. One end of the range represents results of all studies; the other end, results of only the higher certainty studies. Both estimates of the treatment effect (that is, estimate A and estimate B) are calculated to reflect the male-female distribution of the population, as well as the differing reliability of the existing estimates for male patients.

**Table 4.3: Example of a Plan to Account for Multiple Cross-study Differences in Four Randomized Studies<sup>a</sup>**

Cross-study difference	Component of three-part plan
Level of certainty	<p>First, plan to develop two estimates, which may be presented separately or as a range.</p> <p>—Estimate A will be based on all 4 studies (3 male, 1 female).</p> <p>—Estimate B will be based on only the 3 higher certainty studies (2 male, 1 female).</p>
Study specifications (male or female subjects)	<p>Second, in calculating each estimate plan to use a model of gender distribution in patient population:</p> <p>—For estimate A, plan to calculate a separate estimate for males (3 studies) and for females (1 study) and then take weighted average, using weight of 0.5 for male estimate and 0.5 for female estimate.</p> <p>—For estimate B, the plan is the same as for A, except that the male estimate is based on only the 2 higher certainty studies.</p>
Study reliability	<p>Third, when calculating the male estimate, plan to weight each male study by the inverse of its variance.<sup>b</sup></p> <p>—The procedures for calculating male estimates A and B are identical, except that all 3 male studies are included for A and only the 2 higher certainty male studies are used for B.</p>

<sup>a</sup>The four randomized studies include 3 higher certainty studies and one lower quality study. One of the higher certainty studies covers female patients only, while the other 3 studies cover males only. The male studies have differing sample sizes.

<sup>b</sup>Weighting each male study by the inverse of its variance assumes that each of these studies is estimating the same population value.



Many such plans are possible, and the choice of any one plan is guided by investigator judgment, especially on the relative import of the various types of differences. For instance, in the example just cited, we did not specify whether the differences in treatment effects observed for male and female patients were very large; nor did we specify the sizes of the differences in treatment effects observed for the different quality male-only randomized studies.

Because different plans and options may yield different results, the investigator is advised to check the uncertainty associated with this step through a sensitivity analysis. That is, we believe the investigator should conduct this step in two or more alternative ways; this would allow estimation of the difference in synthesis results that alternative plans might produce. This type of sensitivity analysis has been recommended for more traditional meta-analyses (L'Abbé, Detsky, and O'Rourke, 1987).

---

## Synthesizing Results Across Design Categories

Having combined study results within the separate design categories, the next step is to combine results across categories. As discussed earlier in this chapter, combining results across design categories involves dealing with two potential problems. The first problem consists of the “empty set” that is crucial when estimating a treatment effect across the full range of patients: the set of patient population groups not covered by any existing randomized study (stratum 1b of table 4.2 on page 88). The second problem consists of differences in estimates across design categories, which may be linked to differences in certainty, in study designs, or in reliability. Successfully dealing with these problems amounts to reaping the benefits of a cross design synthesis that is based on studies with complementary strengths and weaknesses.

---

## Projecting to an “Empty” Design Category

Projection to patients not covered by randomized studies, using data from all other strata, would be consistent with Rubin's (1990a) approach. The specific form of the projection depends on the assumptions that the investigator makes. Using the synthesis framework shown in table 4.2 on page 88, one method of making such a projection would be to assume that stratum 1a results are to stratum 1b results as stratum 2a is to stratum 2b.<sup>16</sup> Since estimates are available for all except stratum 1b, it is easy to “solve for” the projection.

---

<sup>16</sup>It is important to note that this assumption is made for study results that (1) have been subjected to secondary adjustment and (2) have been combined within each stratum so as to maximize reliability, take account of differences in certainty, and so forth. (This approach is related to that used by Colditz, Miller, and Mosteller, 1988).

Suppose, for example, that existing randomized studies are limited to white patients. One might make the simple assumption that the unknown randomized study results for black patients bear the same relationship to the randomized study results for whites that the observed data base results for blacks bear to data base results for whites. Thus, for example, if the treatment effect that the data base estimates for blacks is 50 percent higher than for whites, the projected randomized study treatment effect for blacks would be 50 percent higher than the observed randomized study treatment effect for whites.<sup>17</sup>

This approach has the advantage of trying to use the maximum available information. Various other options for projection are possible, of course, using alternative assumptions. For example, the relatively common practice of generalizing results from randomized studies covering certain patient groups to other, uncovered patient groups is based on the assumption that results for stratum 1a and stratum 1b are identical. In fact, this assumption has been encouraged by some researchers—so long as there is no reason to believe that the uncovered groups would respond differently than those who participated.<sup>18</sup> In our view, an indication of whether or not this is so is afforded by the comparison of stratum 2a results to stratum 2b results.

## Dealing With Differences in Estimates Across Design Categories

The sizes and patterns of the differences between the estimates for the various strata, the reliability of these estimates and their certainty should all guide the investigator in choosing among options for synthesizing estimates across design categories. One option is to present each stratum estimate separately. Another is to present only estimates from certain strata (such as those deemed to be of higher quality). Many other options derive from the various methods that were described earlier as ways of combining results within design categories. To briefly summarize these methods, they include:

- taking an unweighted average of treatment effects estimated for two or more categories (strata);

<sup>17</sup>Of course, hidden artifacts biasing the stratum 1a or the stratum 1b estimate (or both) would, in turn, bias the projection. Ideally, where the investigator is conscious that uncertainty must be associated with the multistudy estimate for a stratum, he or she would develop a range for that stratum, perhaps based on a sensitivity analysis. The projection would then involve use of a range for one or more strata.

<sup>18</sup>Indeed, the English meta-analyst I. Chalmers (1989) maintains that those who would challenge this practice in any one instance should provide evidence that patients not included in a randomized study do respond differently to the treatment in question than those who were included in the study.



- taking a weighted average of treatment effects estimated for two or more categories;
- using other models or projections to combine results from two or more categories;
- using stratification or ranges to retain differences in estimates of the treatment effect for different strata; and
- developing an overall plan to combine various methods strategically.

The investigator's choice of a particular option (or decision to develop a plan for using a set of options) will depend upon the specific pattern of cross design estimates. This includes (1) whether results converge or diverge across design strata and (2) whether there are differences in the levels of certainty associated with the adjusted multistudy estimates derived from the different strata.

For example, referring back to table 4.2 on page 88, many investigators might decide that stratum 1a results were of better quality than stratum 1b results. Such investigators might decide to use the stratum 1a results alone for patient groups covered by randomized studies. Then, for groups not covered by randomized studies, these investigators would use the stratum 1b projection (based on all other strata) by itself.

Alternatively, however, suppose that an investigator believed that stratum 1a and stratum 2a had produced estimates of a similar level of quality, all things considered. If these two estimates for patient groups covered by randomized studies differed from each other, the investigator might present a range based on the two different estimates. Then, for patient groups not covered by randomized studies, after deriving a projection for stratum 1b, the investigator could again present a range of values—this time based on the projection for stratum 1b and the estimate for stratum 2b.

Or suppose that stratum 1a results and stratum 2a results were nearly identical, but that these differed considerably from the estimate for patient groups not covered by randomized studies (stratum 2b). It might be

appropriate to average only 1a and 2a.<sup>19</sup> The estimate for stratum 2b would then be presented separately.

## **Summary of Tasks 3 and 4: Adjusting and Combining Studies**

To summarize, having assessed generalizability problems in existing randomized studies and comparison bias in data base analyses, the cross design investigator is faced with the challenges of adjusting and combining studies. These challenges include dealing with known biases in individual studies (identified in assessment) and with cross-study differences, which include major differences in study designs and in the patient populations covered. In adjusting and combining studies, the investigator must successfully meet these challenges. The tasks, steps, and methodological options for doing so are summarized in table 4.4. Taken together with the assessment methods presented in earlier chapters, these constitute a first-cut methodology for an investigator to use in conducting a cross design synthesis. However, many refinements are still to be developed. In particular, we believe that future developmental work should target new procedures to minimize the role of investigator subjectivity in determining the conclusions of the synthesis.

<sup>19</sup>Averaging multistudy results from stratum 1a and stratum 2a (rather than using results from 1a alone) is appropriate when: (1) the results from the two strata are similar, (2) the existing randomized studies are few in number or marked by small sample size, and (3) the data base analyses are marked by minimal levels of uncertainty regarding comparison bias. Other times, for patient groups covered by randomized studies, stratum 1a results alone may be deemed superior. Such judgments must be made by the investigator.



**Chapter 4**  
**Methods for Adjusting and Combining**  
**Results of Randomized Studies and Data**  
**Base Analyses (Tasks 3 and 4)**

**Table 4.4: Set of Strategies for Adjusting and Combining Diverse Studies**

<b>Major steps (from table 4.1)</b>	<b>Examples of methodological options for each step</b>
1. Adjust each randomized study's treatment effect.	Standardize results to correct for overrepresentation or underrepresentation (Fleiss, 1973; Deming, 1964).
2. Adjust results from each data base analysis.	Conduct secondary analysis of data base. Other options include use of a ratio (see Eddy, Hasselblad, and Schacter, 1989).
3. Create a framework to organize results.	Stratify studies by type of design (Light and Pillemer, 1984) and by coverage of patient subgroups (Himel et al., 1986). Match data base patients to those covered in randomized studies (Hlatky, 1991); identify those remaining data base patients not covered in randomized studies.
4. Within each stratum of the framework, combine estimates of the treatment effect, adjusting (or otherwise accounting) for differences in quality, in studies' population coverage or procedures, and in reliability. <sup>a</sup>	Use models that account for differences (Eddy, Hasselblad, and Schacter, 1989); take a weighted average with weights defined by the inverse of variances (Hedges and Olkin, 1985); weight studies according to level of certainty; zero weights possible (Rosenthal, 1984); project to ideal study (Rubin, 1990a).  Use a range of estimates or separate estimates for substrata (defined by different study populations, or by certainty or quality level).  Develop plan (see table 4.3; see also GAO, 1989b) for combining these options.
5. Synthesize estimates across design categories (that is, across strata of framework from step 3). As appropriate:	
Provide an estimate for the "null stratum" (stratum 1b in table 4.2).	Project to the empty stratum, using results from other strata (see Rubin, 1990a; Colditz, Miller, and Mosteller, 1988).
Combine estimates across design categories, adjusting (or otherwise accounting) for differences in estimates linked to quality design, or reliability. <sup>a</sup>	Same options as for step 4 above (including use of zero weights as deemed appropriate by investigator).

<sup>a</sup>By differences in quality, we mean differences in the level of certainty that the cross design investigator associates with (1) the adjusted results of each study or (2) the multistudy estimate from each design category.

# List of Experts

Consultants contributed to this study during different phases of the work. Some reviewed our approach at the outset while others reviewed a draft as the study neared completion. Some were interested in the “cross design synthesis” as an overall approach, while others were helpful primarily for a single chapter or for a particular aspect of the discussion that appears in several chapters.

David Cordray, Ph.D., Department of Human Resources, Vanderbilt University, Nashville, Tenn.

Rebecca Gelman, Ph.D., Dana-Farber Cancer Institute, Harvard University, Boston, Mass.

Henry Krakauer, M.D., Uniformed Services University of the Health Sciences, Bethesda, Md.

William Kruskal, Ph.D., Department of Statistics, University of Chicago, Chicago, Ill.

Richard Light, Ph.D., John F. Kennedy School of Government, Harvard University, Cambridge, Mass.

Mark Lipsey, Ph.D., Psychology Department, Claremont University Graduate School, Claremont, Calif.

Clement McDonald, M.D., Regenstrief Institute, Indiana University School of Medicine, Indianapolis, Ind.

Robert Moffitt, Ph.D., Department of Economics, Brown University, Providence, R.I.

Frederick Mosteller, Ph.D., Technology Assessment Group, Harvard School of Public Health, Boston, Mass.

David Pryor, M.D., Duke University Medical Center, Durham, N.C.

Donald Rubin, Ph.D., Department of Statistics, Harvard University, Cambridge, Mass.

Henry Sacks, M.D., Mount Sinai Medical Center, New York, N.Y.

Lee Sechrest, Ph.D., Department of Psychology, University of Arizona, Tucson, Ariz.



# Major Contributors to This Report

---

## Program Evaluation and Methodology Division

George Silberman, Assistant Director  
Judith A. Droitcour, Project Manager  
Elizabeth W. Scullin, Reports Analyst

# Bibliography

---

Altman, Douglas G., and Caroline J. Doré. "Randomisation and Baseline Comparisons in Clinical Trials," Lancet, 335:149-53, 1990.

Andersen, Tavs Folmer, Henrik Bronnum-Hansen, Torben Sejr, and Christian Roepstorff. "Elevated Mortality Following Transurethral Resection of the Prostate for Benign Hypertrophy! But Why?" Medical Care, 28:870-79, 1990.

Angrist, Joshua D. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," American Economic Review, 80:313-36, 1990.

Armitage, P., and M. Hills. "The Two-Period Cross-over Trial," Statistician, 31:119-31, 1982.

Atlay, R.D., A.R. Weekes, G.D. Entwistle, and D.J. Parkinson, "Treating Heartburn in Pregnancy: A Comparison of Acid and Alkali Mixtures," British Medical Journal, 2:919-20, 1978.

Badwe, R.A., W.M. Gregory, M.A. Chaudary, M.A. Richards, A.E. Bentley, R.D. Rubens, I.S. Fentiman. "Timing of Surgery During Menstrual Cycle and Survival of Premenopausal Women With Operable Breast Cancer," Lancet, 337:1261-64, 1991.

Barnett, H.J.M., D. Sackett, D.W. Taylor, B. Haynes, S.J. Peerless, I. Meissner, V. Hachinski, and A. Fox. "Are the Results of the Extracranial-Intracranial Bypass Trial Generalizable?" New England Journal of Medicine, 316:820-24, 1987.

Barnow, Burt S., Glen G. Cain, and Arthur S. Goldberger. "Issues in the Analysis of Selectivity Bias." In Ernst W. Stromsdorfer and George Farkas (eds.), Evaluation Studies Review Annual, Vol. 5. Beverly Hills: Sage, 1980.

Becker, Betsy Jane. "Synthesizing Standardized Mean-change Measures," British Journal of Mathematical and Statistical Psychology, 41:257-78, 1988.

Beecher, Henry K. "The Powerful Placebo," Journal of the American Medical Association, 159:1602-06, 1955.



---

Begg, Colin B., P.B. McGlave, and Louise Pilote. "Bone-marrow Transplantation Versus Chemotherapy in Acute Nonlymphocytic Leukemia: A Meta-analytic Review," European Journal of Cancer and Clinical Oncology, 25:1519-23, 1989.

Begg, Colin B., and Louise Pilote. "A Model for Incorporating Historical Controls Into a Meta-Analysis," Biometrics, 47:899-906, 1991.

Bentler, Peter M. "Latent Variable Structural Models for Separating Specific From General Effects." In Lee Sechrest et al. (eds.), Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data. Rockville, Md.: Agency for Health Care Policy and Research, 1990.

Bentler, Peter M., and Douglas G. Bonett. "Significance Tests and Goodness of Fit in the Analysis of Covariance Structures," Psychological Bulletin, 88:588-606, 1980.

Berk, Richard A. "An Introduction to Sample Selection Bias in Sociological Data," American Sociological Review, 48:386-98, 1983.

Bill, J., R. Anderson, J. O'Fallon, et al. "Development of a Computerized Cancer Data Management System at the Mayo Clinic," International Journal of Biomedical Computing, 9:477, 1978.

Boruch, Robert F. "Comparative Aspects of Randomised Experiments for Planning and Evaluation." In Martin Bulmer (ed.), Social Science Research and Government. Cambridge: Cambridge University Press, 1987.

Boruch, Robert F., David S. Cordray, and Paul M. Wortman. "Secondary Analysis: Why, How, and When." In Robert F. Boruch, Paul M. Wortman, David S. Cordray and Associates (eds.), Reanalyzing Program Evaluations. San Francisco: Jossey-Bass, 1981.

Bracken, Michael, Murray Enkin, Hubert Campbell, and Iain Chalmers. "Symptoms in Pregnancy: Nausea and Vomiting, Heartburn, Constipation, and Leg Cramps." In Iain Chalmers, Murray Enkin, and Marc J.N.C. Keirse (eds.), Effective Care in Pregnancy and Childbirth. New York: Oxford University Press, 1989.

Byar, David P. "Why Data Bases Should Not Replace Randomized Clinical Trials," Biometrics, 36:337-42, 1980.

---

Byar, David P., David A. Schoenfeld, Sylvan B. Green, et al. "Design Considerations for AIDS Trials," New England Journal of Medicine, 323:1343-48, 1990.

Cain, G. "Regression and Selection Models To Improve Nonexperimental Comparisons." In C.A. Bennett and A.A. Lumsdaine (eds.), Evaluation and Experiments. New York: Academic Press, 1975.

Califf, Robert M., David B. Pryor, and Joseph C. Greenfield. "Beyond Randomized Clinical Trials: Applying Clinical Experience in the Treatment of Patients With Coronary Artery Disease," Circulation, 74:1191-94, 1986.

Campbell, Donald T., and A.E. Erlebacher. "How Regression Artifacts in Quasi-Experimental Evaluations Can Mistakenly Make Compensatory Education Look Harmful." In J. Hellmuth (ed.), Compensatory, Education: A National Debate, Vol. 3: The Disadvantaged Child. New York: Brunner/Mazel, 1970.

Campbell, Donald T., and Julian C. Stanley. Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally College Publishing Co., 1963.

Chalmers, Iain. "Evaluating the Effects of Care During Pregnancy and Childbirth." In Iain Chalmers, Murray Enkin, and Marc J.N.C. Keirse (eds.), Effective Care in Pregnancy and Childbirth. New York: Oxford University Press, 1989.

Chalmers, Iain, Murray Enkin, and Marc J.N.C. Keirse (eds.). Effective Care in Pregnancy and Childbirth. New York: Oxford University Press, 1989.

Chalmers, Thomas C., Harry Smith, Jr., Bradley Blackburn, Bernard Silverman, Biruta Schroeder, Dinah Reitman, and Alexander Ambroz. "A Method for Assessing the Quality of a Randomized Control Trial," Controlled Clinical Trials, 2:31-49, 1981.

Chang, I.M., Rebecca S. Gelman, and M. Pagano. "Corrected Group Prognostic Curves and Summary Statistics," Journal of Chronic Diseases, 35:669-75, 1982.



---

Charlson, M.E., P. Pompei, K.L. Ales, C.R. MacKensie. "A New Method of Classifying Prognostic Comorbidity in Longitudinal Studies: Development and Validation," Journal of Chronic Diseases, 40:373-83, 1987.

Chassin, Mark R., Jacqueline Kosecoff, R.E. Park, Constance M. Winslow, Katherine L. Kahn, Nancy J. Merrick, Joan Keesey, Arlene Fink, David Solomon, and Robert H. Brook. "Does Inappropriate Use Explain Geographic Variations in the Use of Health Care Services? A Study of Three Procedures," Journal of the American Medical Association, 258:2533-37, 1987.

Cochran, William G. Sampling Techniques, 2nd and 3rd eds. New York: Wiley, 1963, 1977.

Cochran, William G. "The Planning of Observational Studies of Human Populations," Journal of the Royal Statistical Society, Series A, 128:234-55, 1965.

Cochran, William G. "The Combination of Estimates From Different Experiments," Biometrics, 10:101-29, 1954.

Cochran, William G. "Problems Arising in the Analysis of a Series of Similar Experiments," Journal of the Royal Statistical Society, Supp. 4:102-18, 1937.

Cochran, William G., and S.P. Carroll. "A Sampling Investigation of the Efficiency of Weighting Inversely as the Estimated Variance," Biometrics, 9:447-59, 1953.

Cochran, William G., and G. Cox. Experimental Designs. New York: Wiley, 1957.

Cochran, William G., and Donald B. Rubin. "Controlling Bias in Observational Studies: A Review," Sankhya: The Indian Journal of Statistics, Series A, 35:417-46, 1973.

Colditz, G., J. Miller, and F. Mosteller. "The Effect of Study Design on Gain in Evaluation of New Treatments in Medicine and Surgery," Drug Information Journal, 22:343-52, 1988.

---

Collins, Rory, Richard Peto, Stephen MacMahon, Patricia Hebert, Nicholas H. Fiebach, Kimberly A. Eberlein, Jon Godwin, Nawab Qizilbash, James O. Taylor, and Charles H. Hennekens. "Blood Pressure, Stroke, and Coronary Heart Disease, Part 2, Short-Term Reductions in Blood Pressure: Overview of Randomized Drug Trials in Their Epidemiological Context," Lancet, 335:827-37, 1990.

Connell, Frederick A., Paula Diehr, L. Gary Hart. "The Use of Large Data Bases in Health Care Studies," Annual Review of Public Health, 8:51-74, 1987.

Cook, Thomas, and Donald T. Campbell. Quasi-Experimentation: Design and Analysis Issues for Field Settings. Chicago: Rand McNally, 1979.

Cordray, David S. "An Assessment From the Policy Perspective." In Kenneth W. Wachter and Miron L. Straf (eds.), The Future of Meta-Analysis. New York: Russell Sage Foundation, 1990a.

Cordray, David S. "Strengthening Causal Interpretations of Nonexperimental Data: The Role of Meta-Analysis." In Lee Sechrest et al. (eds.), Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data. Rockville, Md.: Agency for Health Care Policy and Research, 1990b.

Cornfield, J., W. Haenszel, E.C. Hammond, A.M. Lillienfeld, M.B. Shimkin, and E.L. Wynder. "Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions," Journal of the National Cancer Institute, 22:173-203, 1959.

Coyle, Susan L., Robert F. Boruch, and Charles F. Turner (eds.). Evaluating AIDS Prevention Programs, Expanded edition. Washington, D.C.: National Academy Press, 1991.

Cox, David R. "Partial Likelihood." Biometrika, 62:269-76, 1975.

Cox, David R. "Regression Models and Life Tables," Journal of the Royal Statistical Society, Series B, 34:187-202, 1972.

Cox, David R. Planning of Experiments. New York: Wiley, 1958.

Cronbach, Lee J. Designing Evaluations of Educational and Social Programs. San Francisco: Jossey-Bass, 1982.



Cronbach, Lee J., D.R. Rogosa, R.E. Floden, and G.G. Price. "Analysis of Covariance in Nonrandomized Experiments: Parameters Affecting Bias." Occasional paper. Berkeley, Calif.: Stanford Evaluation Consortium, 1977.

Davis, K. "The Comprehensive Cohort Study: The Use of Registry Data to Confirm and Extend a Randomized Trial," Recent Results in Cancer Research, Vol. 111. Berlin-Heidelberg: Springer-Verlag, 1988.

Davis, Scott, Peter W. Wright, Susan F. Schulman, Lucius D. Hill, Roland D. Pinkham, Lloyd P. Johnson, Thomas W. Jones, Howard B. Kellogg, Jr., Hubert M. Radke, Wesley W. Sikkema, Philip C. Jolly, and Samuel P. Hammar. "Participants in Prospective, Randomized Clinical Trials for Resected Non-Small Cell Lung Cancer Have Improved Survival Compared With Nonparticipants in Such Trials," Cancer, 56:1710-18, 1985.

Deming, W. Edwards. Statistical Adjustment of Data. New York: Dover, 1964.

Deming, W. Edwards. Sample Design in Business Research. New York: Wiley, 1960.

Deming, W. Edwards. Some Theory of Sampling. New York: Dover, 1960.

Demlo, Linda K. "Measuring Health Care Effectiveness: Research and Policy Implications," International Journal of Technology Assessment in Health Care, 6:288-94, 1990.

Devine, Elizabeth, and Thomas D. Cook. "A Meta-Analytic Analysis of Effects of Psychoeducational Interventions on Length of Postsurgical Hospital Stay," Nursing Research, 32:267-74, 1983.

DuBois, Philip H. Multivariate Correlational Analysis. New York: Harper and Row, 1957.

Eddy, David M. "The Confidence Profile Method: A Bayesian Method for Assessing Health Technologies," Operations Research, 37:210-28, 1989.

Eddy, David M., Victor Hasselblad, William McGivney, and William Hendee. "The Value of Mammography Screening in Women Under Age 50," Journal of the American Medical Association, 259:1512-19, 1988.

---

Eddy, David M., Vic Hasselblad, and Ross Shachter. The Statistical Synthesis of Evidence: Meta-Analysis by the Confidence Profile Method. Report issued by the Center for Health Policy Research and Education, Duke University, and by the Department of Engineering-Economic Systems, Stanford University, 1989.

Edlund, Matthew J., Thomas J. Craig, and Mary Ann Richardson. "Informed Consent as a Form of Volunteer Bias," American Journal of Psychiatry, 142:624-27, 1985.

Ellenberg, Susan S. "Meta-Analysis: The Quantitative Approach to Research Review," Seminars in Oncology, 15:472-81, 1988.

Ellwood, Paul M. "A Technology of Patient Experience," New England Journal of Medicine, 318:1549-56, 1988.

Fisher, Bernard, and Carol Redmond. "Letter to the Editor," New England Journal of Medicine, 321:472, 1989.

Fisher, Bernard, Nelson Slack, Donna Katrych, and Norman Wolmark. "Ten Year Follow-Up Results of Patients With Carcinoma of the Breast in a Co-operative Clinical Trial Evaluating Surgical Adjuvant Chemotherapy," Surgery, Gynecology & Obstetrics, 140:528-34, 1975.

Fisher, Ronald A. Statistical Methods for Research Workers, 1st ed. Edinburgh: Oliver and Boyd, 1925.

Fisher, Ronald A. The Design of Experiments. Edinburgh: Oliver and Boyd, 1935.

Fleiss, Joseph L. Statistical Methods for Rates and Proportions. New York: Wiley, 1973.

Francis, Thomas, Jr., Robert F. Korns, Robert B. Voight, Morton Boisen, Fay M. Hemphill, John A. Napier, and Eva Tolchinsky. An Evaluation of the 1954 Poliomyelitis Vaccine Trials: Summary Report. Ann Arbor, Mich.: The Poliomyelitis Vaccine Evaluation Center, University of Michigan, 1955.

GAO. See U.S. General Accounting Office.



Garceau, A.J., R.M. Donaldson, E.T. O'Hara, et al. "A Controlled Trial of Prophylactic Portacaval-shunt Surgery," New England Journal of Medicine, 270:496-500, 1964.

Gehan, Edmund. "The Evaluation of Therapies: Historical Control Studies," Statistics in Medicine, 3:315-24, 1984.

Gillings, Dennis, James Grizzle, Gary Koch, Karl Rickels, Ingrid Amara, Mary Donelan, Stephen Hardiman, Ralph Nash, William Sollecito, and William Stager. "Pooling 12 Nomifensine Studies for Efficacy Generalizability," Journal of Clinical Psychiatry, 45:78-84, 1984.

Glass, Gene V. Book review of "The Future of Meta-Analysis," Journal of the American Statistical Association, 86:1141, 1991.

Glass, Gene V. "In Defense of Generalization," The Behavioral and Brain Sciences, 3:394-95, 1978.

Glass, Gene V. "Primary, Secondary, and Meta-analysis of Research," Educational Researcher, 6:3-8, 1976.

Glass, Gene V., Barry McGaw, and Mary Lee Smith. Meta-Analysis in Social Research. Beverly Hills: Sage, 1981.

Goldberger, A.S., and O.D. Duncan (eds.). Structural Equation Models in the Social Sciences. New York: Seminar, 1973.

Greenfield, Sheldon. "The State of Outcome Research: Are We On Target?" New England Journal of Medicine, 320:1142-43, 1989.

Hansen, M.H., and W.N. Hurwitz. "The Problem of Nonresponse in Sample Surveys," Journal of the American Statistical Association, 41:517-29, 1946.

Harrison, Harriett H., and Julie Morgan. "Quality Control of Screening Procedures in the Multiple Risk Factor Intervention Trial," Controlled Clinical Trials, 7:91S-108S, 1986.

Heckman, James J., and V. Joseph Hotz. "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," Journal of the American Statistical Association, 84:862-74, 1989a.

---

Heckman, James J., and V. Joseph Hotz, "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," Working Paper No. 2861. Cambridge, Mass.: National Bureau of Economic Research, Inc., 1989b.

Heckman, James J., and V. Joseph Hotz. "Rejoinder," Journal of the American Statistical Association, 84:878-80, 1989c.

Hedges, Larry V. "Directions for Future Methodology." In Kenneth W. Wachter and Miron L. Straf (eds.), The Future of Meta-Analysis. New York: Russell Sage Foundation, 1990.

Hedges, Larry V. "Advances in Statistical Methods for Meta-Analysis." In William H. Yeaton and Paul M. Wortman, Issues in Data Synthesis: New Directions for Program Evaluation. San Francisco: Jossey-Bass, 1984.

Hedges, Larry V., and Ingram Olkin. Statistical Methods for Meta-Analysis. New York: Academic Press, 1985.

Himel, Harvey N., Alessandro Liberati, Richard Gelber, and Thomas C. Chalmers. "Adjuvant Chemotherapy for Breast Cancer: A Pooled Estimate Based on Published Randomized Control Trials," Journal of the American Medical Association, 256:1148-59, 1986.

Hlatky, Mark A. "Using Databases to Evaluate Therapy," Statistics in Medicine, 10:647-52, 1991.

Hlatky, Mark A., Robert M. Califf, Frank E. Harrell, Jr., Kerry L. Lee, Daniel B. Mark, and David B. Pryor. "Comparison of Predictions Based on Observational Data With the Results of Randomized Controlled Clinical Trials of Coronary Artery Bypass Surgery," Journal of the American College of Cardiology, 11:237-45, 1988.

Holland, Paul W. "It's Very Clear" (Comment), Journal of the American Statistical Association, 84:875-77, 1989.

Holland, Paul W. "Statistics and Causal Inference," Journal of the American Statistical Association, 81:945-70, 1986.

Holland, Paul W., and Donald B. Rubin. "On Lord's Paradox." In H. Wainer and S. Messick (eds.), Principals of Modern Psychological Measurement: Festschrift for Frederick M. Lord. Hillsdale, N.J.: Erlbaum, 1983.



---

Hornbrook, Mark C. "Techniques for Assessing Hospital Case Mix," Annual Review of Public Health, 6:295-324, 1985.

Hovell, Melbourne F. "The Experimental Evidence for Weight-Loss Treatment of Essential Hypertension: A Critical Review," American Journal of Public Health, 72:359-68, 1982.

Hrushesky, William A., Avrum Z. Bluming, and Scott A. Gruber. "Menstrual Influence on Surgical Cure of Breast Cancer" (Letter to the Editor), Lancet, 335:984, 1990.

Hrushesky, William A., Avrum Z. Bluming, Scott A. Gruber, and Robert B. Sothorn. "Menstrual Influence on Surgical Cure of Breast Cancer," Lancet, II(8669):949-52, 1989.

Hunter, John E., and Frank L. Schmidt. Methods of Meta-Analysis. Newbury Park, Calif.: Sage Publications, 1990.

Hyman, Herbert. Survey Design and Analysis. New York: Free Press, 1955.

Jackson, Gregg B. "Methods for Integrative Reviews," Review of Educational Research, 50:438-60, 1980. (Reprinted in Richard J. Light (ed.), Evaluation Studies Review Annual, Vol. 8. Beverly Hills: Sage, 1983.)

Jenicek, Milos. "Meta-Analysis in Medicine: Where We Are and Where We Want to Go," Journal of Clinical Epidemiology, 42:35-44, 1989.

Joreskog, K.G. "Structural Equation Models in the Social Sciences: Specification, Estimation and Testing." In P.R. Krishnaiah (ed.), Applications of Statistics. Amsterdam: North-Holland, 1977.

Kalton, Graham. Compensating for Missing Data. Ann Arbor, Mich.: Institute for Social Research, University of Michigan, 1983.

Kaplan, Robert M., and Charles C. Berry. "Adjusting for Confounding Variables." In Lee Sechrest et al. (eds.), Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data. Rockville, Md.: Agency for Health Care Policy and Research, 1990.

Kirsch, Irving, and Lynne J. Weixel. "Double-Blind Versus Deceptive Administration of a Placebo," Behavioral Neuroscience, 102:319-23, 1988.

---

Kish, Leslie. Statistical Design for Research. New York: Wiley, 1987.

Kleinbaum, David G., Lawrence L. Kupper, and Hal Morgenstern. Epidemiologic Research: Principles and Quantitative Methods. New York: Van Nostrand Reinhold, 1982.

Krakauer, Henry. "Assessment of Alternative Technologies for the Treatment of End-Stage Renal Disease," Israel Journal of Medical Sciences, 22:245-59, 1986.

Krakauer, Henry, and R. Clifton Bailey. "Epidemiologic Oversight of the Medical Care Provided to Medicare Beneficiaries," Statistics in Medicine, 10:521-40, 1991.

Kramer, Michael S., and Stanley H. Shapiro. "Scientific Challenges in the Application of Randomized Trials," Journal of the American Medical Association, 252:2739-45, 1984.

Kruskal, William and Frederick Mosteller. "Ideas of Representative Sampling." In D. Fiske (ed.), New Directions for Methodology of Social and Behavioral Science: Problems With Language Imprecision. San Francisco: Jossey-Bass, 1981.

Kurland, L.T., and C.A. Molgaard. "The Patient Record in Epidemiology," Scientific American, 245:54, 1981.

L'Abbé, Kristan A., Allan S. Detsky, and Keith O'Rourke. "Meta-Analysis in Clinical Research," Annals of Internal Medicine, 107:224-33, 1987.

Laszlo, John, John C. Bailar III, and Frederick Mosteller. "Registers and Data Bases." In Frederick Mosteller et al., Assessing Medical Technologies, pp. 101-09. Washington, D.C.: National Academy Press, 1985.

Lavori, Philip W., Thomas A. Louis, John C. Bailar III, and Marcia Polansky. "Designs for Experiments: Parallel Comparisons of Treatment," New England Journal of Medicine, 309:1291-99, 1983.

Lazarsfeld, Paul, Ann Pasanella, and Morris Rosenberg (eds.). Continuities in the Language of Social Research. New York: Free Press, 1972.



Leveno, Kenneth J., F. Gary Cunningham, Sheryl Nelson, Micki Roark, M. Lynne Williams, David Guzick, Sharon Dowling, Charles R. Rosenfeld, and Ann Buckley. "A Prospective Comparison of Selective and Universal Electronic Fetal Monitoring in 34,995 Pregnancies," New England Journal of Medicine, 315:615-19, 1986.

Liberati, Alessandro, Andre L. Blum, Giovanni Apolone, and Antonio Nicolucci. "Basic Principles for Use and Interpretation of Epidemiologic Data." In Thomas Chalmers et al. (eds.), Data Analysis for Clinical Medicine: The Quantitative Approach to Patient Care in Gastroenterology. New York: International University Press, 1988.

Liberati, Alessandro, Harvey N. Himel, and Thomas C. Chalmers. "A Quality Assessment of Randomized Control Trials of Primary Treatment of Breast Cancer," Journal of Clinical Oncology, 4:942-51, 1986.

Lichtman, Stuart M., and Daniel R. Budman. Letter to the Editor, New England Journal of Medicine, 321:470, 1989.

Light, Richard J. "Six Evaluation Issues That Synthesis Can Resolve Better Than Single Studies." In William Yeaton and Paul Wortman (eds.), Issues in Data Synthesis: New Directions for Program Evaluation. San Francisco: Jossey-Bass, 1984.

Light, Richard J., and David B. Pillemer. Summing Up: The Science of Reviewing Research. Cambridge, Mass.: Harvard University Press, 1984.

Lind, J. A Treatise of the Scurvy. Edinburgh: Sands Murray and Cochran, 1753.

Lipsey, Mark. "Juvenile Delinquency Treatment: A Meta-analytic Inquiry Into the Variability of Effects." In Thomas D. Cook, Harris Cooper, David S. Cordray, Heidi Hartmann, Larry V. Hedges, Richard J. Light, Thomas A. Louis, and Frederick Mosteller (eds.), Meta-Analysis for Explanation: A Casebook. New York: Russell Sage Foundation, 1992.

Longnecker, Matthew, Jesse A. Berlin, Michele Orza, and Thomas C. Chalmers. "A Meta-analysis of Alcohol Consumption in Relation to Risk of Breast Cancer," Journal of the American Medical Association, 260:652-56, 1988.

Louis, P.C.A. Recherches sur les Effets de la Saignée. Paris: De Mignaret, 1835.

Louis, P.C.A. Essay on Clinical Instruction (translated by P. Martin). London: S. Highley, 1834.

Louis, Thomas A., Harvey V. Fineberg, and Frederick Mosteller. "Findings for Public Health From Meta-Analyses," Annual Review of Public Health, 6:1-20, 1985.

MacMahon, Stephen, Richard Peto, Jeffrey Cutler, Rory Collins, Paul Sorlie, James Neaton, Robert Abbott, Jon Godwin, Alan Dyer, and Jeremiah Stamler. "Blood Pressure, Stroke, and Coronary Heart Disease; Part 1, Prolonged Differences in Blood Pressure: Prospective Observational Studies Corrected for the Regression Dilution Bias," Lancet, 335:765-74, 1990.

McDonald, Clement, and Siu Hui. "The Analysis of Humongous Databases: Problems and Promises," Statistics in Medicine, 10:511-18, 1991.

McPeck, Bucknam. "Inference, Generalizability, and a Major Change in Anesthetic Practice," Anesthesiology, 66:723-24, 1987.

Medical Research Council. "Clinical Trials of Antihistaminic Drugs in the Prevention and Treatment of the Common Cold," British Medical Journal, ii:425-29, 1950.

Medical Research Council. "Streptomycin Treatment of Pulmonary Tuberculosis," British Medical Journal, ii:769-82, 1948.

Meier, Paul. "The Biggest Public Health Experiment Ever: The 1954 Field Trial of the Salk Poliomyelitis Vaccine." In Judith M. Tanur, Frederick Mosteller, William Kruskal, et al. (eds.), Statistics: A Guide to the Unknown. San Francisco: Holden-Day, 1972.

Merigan, Thomas. "You Can Teach an Old Dog New Tricks: How AIDS Trials Are Pioneering New Strategies," New England Journal of Medicine, 323:1341-43, 1990.

Miké, Valerie. "Clinical Studies in Cancer: A Historical Perspective." In Valerie Miké and Kenneth E. Stanley, Statistics in Medical Research: Methods and Issues With Applications in Cancer Research. New York: Wiley, 1982.



---

Mishel, Merle H. "Confounding Variables." In Lee Sechrest et al. (eds.), Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data. Rockville, Md.: Agency for Health Care Policy and Research, 1990.

Moffitt, Robert. "Program Evaluation With Nonexperimental Data," Evaluation Review, 15(3):291-314, 1991.

Moffitt, Robert. "Comment," Journal of the American Statistical Association, 84:877-78, 1989.

Moon, Thomas E. "Interpretation of Cancer Prevention Trials," Preventive Medicine, 18:721-31, 1989.

Moon, Thomas E., Stephen E. Jones, Gianni Bonadonna, Pinuccia Valagussa, Trevor Powles, Aman Buzdar, and Eleanor Montague. "Development and Use of a Natural History Data Base of Breast Cancer Studies," American Journal of Clinical Oncology (CCT), 10:396-403, 1987.

Moses, Lincoln E. "Innovative Methodologies for Research Using Databases," Statistics in Medicine, 10:629-33, 1991.

Mosteller, Frederick. "Improving Research Methodology: An Overview." In Lee Sechrest et al. (eds.), Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data. Rockville, Md.: Agency for Health Care Policy and Research, 1990a.

Mosteller, Frederick. "Summing Up." In Kenneth W. Wachter and Miron L. Straf (eds.), The Future of Meta-Analysis. New York: Russell Sage Foundation, 1990b.

Mosteller, Frederick, et al. Assessing Medical Technologies, Chapter 3: "Methods of Technology Assessment." Washington, D.C.: National Academy Press, 1985.

Neaton, James D., Richard H. Grimm, Jr., and Jeffrey A. Cutler. "Recruitment of Participants for the Multiple Risk Factor Intervention Trial (MRFIT)," Controlled Clinical Trials, 8:41S-53S, 1987.

---

Neyman, Jerzy Splawa-. "On the Application of Probability Theory to Agricultural Experiments, Essay on Principles, Section 9," Statistical Science, 5:465-80, 1990. (Translated and edited by D.M. Dabrowska and T.P. Speed from the Polish original, which appeared in Roczniki Nauk Rolniczych Tom (Annals of Agricultural Science) X:1-51, 1923.)

Pedhazur, Elazar J. Multiple Regression in Behavioral Research: Explanation and Prediction (2nd ed.). New York: Holt Rinehart Winston, 1982.

Peto, R. "Why Do We Need Systematic Overviews of Randomized Trials?" Statistics in Medicine, 6:233-40, 1987.

Pocock, Stuart. Clinical Trials: A Practical Approach. New York: Wiley, 1983.

Politz, Alfred, and Willard Simmons. "An Attempt To Get the Not-at-Homes Into the Sample Without Callbacks," Journal of the American Statistical Association, 44:9-31, 1949.

Pryor, David B., Robert M. Califf, Frank E. Harrell, Jr., Mark A. Hlatky, Kerry L. Lee, Daniel B. Mark, and Robert A. Rosati. "Clinical Data Bases: Accomplishments and Unrealized Potential," Medical Care, 23(5):623-47, 1985.

Pryor, David B., Frank E. Harrell, Jr., Kerry L. Lee, Robert M. Califf, and Robert A. Rosati. "Estimating the Likelihood of Significant Coronary Artery Disease," American Journal of Medicine, 75:771-80, 1983.

Rao, Poduri S.R.S. "Cochran's Contributions to Variance Components Models for Combining Estimates." In Poduri S.R.S. Rao and Joseph Sedransk (eds.), W.G. Cochran's Impact on Statistics. New York: Wiley, 1984.

Reichardt, Charles S. "The Statistical Analysis of Data From Nonequivalent Group Designs." In Thomas Cook and Donald Campbell, Quasi-Experimentation: Design & Analysis Issues for Field Settings. Chicago: Rand McNally, 1979.

Remington, Richard D. "Potential Impact of Exclusion Criteria on Results of Hypertension Trials," Hypertension, Supp. I, 13:I-66—I-68, 1989.



---

Rindskopf, David. "New Developments in Selection Modeling for Quasi-Experimentation." In William M. K. Trochim (ed.), Advances in Quasi-Experimental Design and Analysis: New Directions in Program Evaluation. San Francisco: Jossey-Bass, 1986.

Rindskopf, David. "Structural Equation Models in Analysis of Nonexperimental Data." In Robert F. Boruch, Paul M. Wortman, David S. Cordray, and Associates, Reanalyzing Program Evaluations. San Francisco: Jossey-Bass, 1981.

Roos, Noralou P., John E. Wennberg, David J. Malenka, Elliott Fisher, Klim McPherson, Tavs Folmer Andersen, Marsha M. Cohen, and Ernest Ramsey. "Mortality and Reoperation After Open and Transurethral Resection of the Prostate for Benign Prostatic Hyperplasia," New England Journal of Medicine, 320:1120-24, 1989.

Roper, William L., William Winkenwerder, Glenn M. Hackbarth, and Henry Krakauer. "Effectiveness in Health Care: An Initiative to Evaluate and Improve Medical Practice," New England Journal of Medicine, 319:1197-1202, 1988.

Rosenbaum, Paul R. "From Association to Causation in Observational Studies: The Role of Tests of Strongly Ignorable Treatment Assignment," Journal of the American Statistical Association, 79:41-48, 1984.

Rosenbaum, Paul R., and Donald B. Rubin. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," Journal of the American Statistical Association, 79:516-24, 1984.

Rosenbaum, Paul R., and Donald B. Rubin. "The Central Role of the Propensity Score in Observational Studies for Causal Effects," Biometrika, 70:41-55, 1983a.

Rosenbaum, Paul R., and Donald B. Rubin. "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study With Binary Outcome," Journal of the Royal Statistical Society, Series B, 45:212-18, 1983b.

Rosenberg, Morris. The Logic of Survey Analysis. New York: Basic Books, 1968.

---

Rosenthal, Robert. "Designing, Analyzing, Interpreting, and Summarizing Placebo Studies." In Leonard White, Bernard Tursky, and Gary E. Schwartz (eds.), Placebo: Theory, Research, and Mechanisms. New York: Guilford, 1985.

Rosenthal, Robert. Meta-Analytic Procedures for Social Research. Beverly Hills: Sage, 1984.

Rossi, Peter H., and Howard E. Freeman. Evaluation: A Systematic Approach (3rd ed.). Beverly Hills: Sage, 1985.

Rowland, Malcolm, Lewis B. Sheiner, and Jean-Louis Steimer (eds.). Variability in Drug Therapy: Description, Estimation, and Control. New York: Raven Press, 1985.

Rubin, Donald B. "Practical Implications of Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism," Biometrics, 47(4):1213-34, December 1991.

Rubin, Donald B. "A New Perspective." In Kenneth W. Wachter and Miron L. Straf (eds.), The Future of Meta-Analysis. New York: Russell Sage Foundation, 1990a.

Rubin, Donald B. "Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies," Statistical Science, 5:472-80, 1990b.

Rubin, Donald B. "William G. Cochran's Contributions to the Design, Analysis, and Evaluation of Observational Studies." In Poduri S.R.S. Rao and Joseph Sedransk (eds.), W.G. Cochran's Impact on Statistics. New York: Wiley, 1984.

Rubin, Donald B. "Bayesian Inference for Causal Effects: The Role of Randomization," Annals of Statistics, 6:34-58, 1978.

Rubin, Donald B. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," Journal of Educational Psychology, 66:688-701, 1974.



Sacks, Henry S., Jayne Berrier, Dinah Reitman, V.A. Ancona-Berk, and Thomas Chalmers. "Meta-Analyses of Randomized Controlled Trials," New England Journal of Medicine, 316:450-55, 1987.

Sacks, Henry S., Thomas C. Chalmers, and Harry Smith, Jr. "Sensitivity and Specificity of Clinical Trials: Randomized v. Historical Controls," Archives of Internal Medicine, 143:753-55, 1983.

Sandmire, H.F. "Whither Electronic Fetal Monitoring?" Obstetrics and Gynecology, 76:1130-34, 1990.

Schatzkin, A., D.Y. Jones, R.N. Hoover, et al. "Alcohol Consumption and Breast Cancer in the Epidemiologic Follow-up Study of the First National Health and Nutrition Examination Survey," New England Journal of Medicine, 316:1169-73, 1987.

Scheffé, Henry. The Analysis of Variance. New York: Wiley, 1959.

Schooler, Nina R. "How Generalizable Are the Results of Clinical Trials?" Psychopharmacology Bulletin, 16:29-31, 1980.

Sechrest, Lee, and Aurelio Jose Figueredo. "Approaches Used in Conducting Outcomes and Effectiveness Research." Paper presented at a conference of the Association for Health Services Research, April 1991.

Sechrest, Lee, and Maureen Hannah. "The Critical Importance of Nonexperimental Data." In Lee Sechrest, et al. (eds.), Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data. Rockville, Md.: Agency for Health Care Policy and Research, 1990.

Senie, Ruby T., Paul Peter Rosen, Philip Rhodes, and Martin L. Lesser. "Timing of Breast Cancer Excision During the Menstrual Cycle Influences Duration of Disease-free Survival," Annals of Internal Medicine, 115:337-42, 1991.

Shadish, William R., Jr., Thomas D. Cook, and Arthur C. Houts. "Quasi-Experimentation in a Critical Multiplist Mode." In William M.K. Trochim (ed.), Advances in Quasi-Experimental Design and Analysis: New Directions for Program Evaluation. San Francisco: Jossey-Bass, 1986.

Speed, T.P. "Introductory Remarks on Neyman (1923)," Statistical Science, 5:463-64, 1990.

Staines, Graham L. "The Strategic Combination Argument." In Werner Leinfellner and Eckehart Köhler (eds.), Developments in the Methodology of Social Science. Boston: Reidel Publishing Co., 1974.

Steimer, Jean-Louis, Alain Mallet, and France Mentré. "Estimating Interindividual Pharmacokinetic Variability." In Malcolm Rowland, Lewis B. Sheiner, and Jean-Louis Steimer (eds.), Variability in Drug Therapy: Description, Estimation, and Control. New York: Raven Press, 1985.

Steinberg, Karen K., Stephen B. Thacker, S. Jay Smith, Donna F. Stroup, Matthew M. Zack, W. Dana Flanders, and Ruth L. Berkelman. "A Meta-analysis of the Effect of Estrogen Replacement Therapy on the Risk of Breast Cancer," Journal of the American Medical Association, 265:1985-90, 1991.

Steinhorn, Sandra, Kenneth J. Kopecky, Max H. Myers, and Charles Ball. "Characteristics of Colon Cancer Patients Reported in Population-Based Tumor Registries and Comprehensive Cancer Centers," Journal of the National Cancer Institute, 70(4):629-34, 1983.

"Student" (William S. Gossett). "The Lanarkshire Milk Experiment." In E.S. Pearson and John Wishart, "Student's" Collected Papers. London: University College, 1942. (Originally published in Biometrika, 23:398, 1931.)

Taylor, Kathryn M., Richard G. Margoese, and Colin L. Soskoline. "Physicians Reasons for Not Entering Eligible Patients in a Randomized Clinical Trial of Surgery for Breast Cancer," New England Journal of Medicine, 310:1363-67, 1984.

Thompson, Simon G., and Stuart J. Pocock. "Can Meta-analyses Be Trusted?" Lancet, 338:1127-30, 1991.

Thompson, Troy L., II, Christopher M. Filley, Wayne D. Mitchell, Kathleen M. Culig, Mary LoVerde, and Richard L. Byyny. "Lack of Efficacy of Hydergine in Patients with Alzheimer's Disease," New England Journal of Medicine, 323:445-48, 1990.

Tierney, William M., and Clement J. McDonald. "Practice Databases and Their Uses in Clinical Research." Statistics in Medicine, 10:541-57, 1991.



U.S. General Accounting Office. Practice Guidelines: The Experience of Medical Specialty Societies (GAO/PEMD-91-11). Washington, D.C.: U.S. General Accounting Office, 1991.

U.S. General Accounting Office. Problems of Implementing the National Institutes of Health Policy on Women in Study Populations (GAO/HRD-90-38). Testimony, June 18, 1990.

U.S. General Accounting Office. Breast Cancer: Patients' Survival (GAO/PEMD-89-9). Washington, D.C.: U.S. General Accounting Office, 1989a.

U.S. General Accounting Office. AIDS Forecasting: Undercount of Cases and Lack of Key Data Weaken Existing Estimates (GAO/PEMD-89-13). Washington, D.C.: U.S. General Accounting Office, 1989b.

U.S. General Accounting Office. Medicare: Improvements Needed in the Identification of Inappropriate Hospital Care (GAO/PEMD-90-7). Washington, D.C.: U.S. General Accounting Office, 1989c.

Wachter, Kenneth W., and Miron L. Straf. "Introduction." In Kenneth W. Wachter and Miron L. Straf (eds.), The Future of Meta-Analysis. New York: Russell Sage, 1990.

Wagner, D.P., W.A. Knaus, and E.A. Draper, "Statistical Validation of a Severity of Illness Measure," American Journal of Public Health 73:878-84, 1983.

Wald, Abraham. "The Fitting of Straight Lines If Both Variables Are Subject to Error," Annals of Mathematical Statistics, 11:284-300, 1940.

Wallace, T. Dudley, and J. Lew Silver. Econometrics: An Introduction. Reading, Mass.: Addison Wesley, 1988.

Weiner, Jonathan P. "Ambulatory Case-Mix Methodologies: Application to Primary Care Research." In Heddy Hibbard, Paul A. Nutting, and Mary L. Grady (eds.), Primary Care Research: Theory and Methods. Rockville, Md.: Agency for Health Care Policy and Research, 1991.

Wennberg, John E., Jean L. Freeman, Roxanne M. Shelton, and Thomas A. Bubolz. "Hospital Use and Mortality Among Medicare Beneficiaries in Boston and New Haven," New England Journal of Medicine, 321:1168-73, 1989.

---

Wennberg, John E., Jean L. Freeman, W.J. Culp. "Are Hospital Services Rationed in New Haven or Over-utilised in Boston?" Lancet, I(8543):1185-9, 1987.

Wennberg, John E., A.G. Mulley, Jr., D. Hanley, et al. "An Assessment of Prostatectomy for Benign Urinary Tract Obstruction: Geographic Variations and the Evaluation of Medical Care Outcomes," Journal of the American Medical Association, 259:3027-30, 1988.

White, B. Alex. "Introduction to Classification and Case Mix in Primary Care." In Heddy Hibbard, Paul A. Nutting, and Mary L. Grady (eds.), Primary Care Research: Theory and Methods. Rockville, Md.: Agency for Health Care Policy and Research, 1991.

Wilhelmsen, Lars, Staffan Ljungberg, Hans Wedel, and Lars Werkö. "A Comparison Between Participants and Non-Participants in a Primary Prevention Trial," Journal of Chronic Diseases, 29:331-39, 1976.

Wilkins, Wallace. "Placebo Controls and Concepts in Chemotherapy and Psychotherapy Research." In Leonard White, Bernard Tursky, and Gary E. Schwartz (eds.), Placebo: Theory, Research, and Mechanisms. New York: Guilford, 1985.

Winslow, Constance Monroe, Jacqueline B. Kosecoff, Mark Chassin, David E. Kanouse, and Robert H. Brook. "The Appropriateness of Performing Coronary Artery Bypass Surgery," Journal of the American Medical Association, 260:505-09, 1988.

Wortman, Paul M., and Fred B. Bryant. "School Desegregation and Black Achievement: An Integrative Review," Sociological Methods and Research, 13:289-324, 1985.

Wortman, Paul M., and William H. Yeaton. "Synthesis of Results in Controlled Trials of Coronary Artery Bypass Graft Surgery." In Richard J. Light (ed.), Evaluation Studies Review Annual, 8:536-51. Beverly Hills: Sage, 1983.

Yancik, R., L.G. Ries, and J.W. Yates. "Breast Cancer in Aging Women: A Population-based Study of Contrasts in Stage, Surgery and Survival," Cancer, 63:976-81, 1989.



---

Yeaton, William. "Causal Power: Strengthening Causal Claims Using No-Difference Findings." In Lee Sechrest et al. (eds.), Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data. Rockville, Md.: Agency for Health Care Policy and Research, 1990.

Yeaton, William H., and Paul M. Wortman. "Medical Technology Assessment: The Evaluation of Coronary Artery Bypass Graft Surgery Using Data Synthesis Techniques," International Journal of Technology Assessment in Health Care, 1:125-46, 1985.

Yusuf, Salim, Richard Simon, and Susan S. Ellenberg. "Preface to the Proceedings of the Workshop on Methodologic Issues in Overviews of Randomized Clinical Trials, May 1986," Statistics in Medicine, 6:217-18, 1987.

Zelen, Marvin, and Rebecca Gelman. "Assessment of Adjuvant Trials in Breast Cancer," National Cancer Institute Monographs, No. 1, 1986.











---

### **Ordering Information**

**The first copy of each GAO report and testimony is free. Additional copies are \$2 each. Orders should be sent to the following address, accompanied by a check or money order made out to the Superintendents of Documents, when necessary. Orders for 100 or more copies to be mailed to a single address are discounted 25 percent.**

**U.S. General Accounting Office  
P.O. Box 6015  
Gaithersburg, MD 20877**

**Orders may also be placed by calling (202)275-6241.**



---

**United States  
General Accounting Office  
Washington, D.C. 20548**

**First-Class Mail  
Postage & Fees Paid  
GAO  
Permit No. G100**

**Official Business  
Penalty for Private Use \$300**

---